A support for understanding medical notes: correcting spelling errors in Italian clinical records

Roger Ferrod[1], Enrico Brunetti[1,2], Luigi Di Caro[1], Chiara Di Francescomarino[3], Mauro Dragoni[3], Chiara Ghidini[3], Renata Marinello[2] and Emilio Sulis[1]

[1]*University of Turin, Torino, Italy*
[2]*City of Health and Science, Torino, Italy*
[3]*Fondazione Bruno Kessler, Treno, Italy*

Abstract

In a context of digitalization and modernization of healthcare, automatic analysis of clinical data plays a leading role in improving the quality of care. Since much of the information lies in an unstructured form within clinical notes, it is necessary to make use of modern Natural Language Processing techniques to extract and build structured knowledge from the data. However, clinical texts pose unique challenges due to the extensive usage of *i)* acronyms, *ii)* non-standard medical jargons and *iii)* typos over technical terms. In this paper, we present a prototype spell-checker specifically designed for medical texts written in Italian.

Keywords

Clinical notes, Natural Language Processing, Spelling correction

1. Introduction

Healthcare is more than ever a priority for every country. Post-COVID-19 healthcare will be characterized by a renewed interest in modernization. Indeed, healthcare is still one of the least digitized industries, although it generates alone 5% of all the data in the world[1]. For these reasons we are witnessing today a revolution that sees data as protagonists, as in the case of Electronic Health Records (EHRs) to systematically collect a wide range of patient data (e.g. medical history, medications, laboratory tests, vital signs etc.). In this context, also self-care processes are currently undergoing modernization and automation, as shown for example in [1, 2, 3].

The digitization of healthcare involves the recent frontiers of the medical internet of things, sensor applications to monitor both human behavior and the environment. In addition, healthcare organizations must pay attention to the role of information systems [4]. The most recent developments in organization management involve automated analysis of information recorded in so-called *event-log* files [5, 6]. Health information systems collect data to leverage digital

AIxIA 2021 SMARTERCARE Workshop, November 29, 2021, Milan, IT
roger.ferrod@unito.it (R. Ferrod); enrico.brunetti@unito.it (E. Brunetti); luigi.dicaro@unito.it (L. D. Caro); dfmchira@fbk.eu (C. D. Francescomarino); dragoni@fbk.eu (M. Dragoni); ghidini@fbk.eu (C. Ghidini); rmarinello@cittadellasalute.to.it (R. Marinello); emilio.sulis@unito.it (E. Sulis)


[1]Source: UBS Group SA, as of June 2020

traces regarding activities, patients, as well as medical notes. In particular, it is relevant to consider both structured and unstructured data, i.e., clinical and textual.

Since a portion of the healthcare data is in textual form, it is increasingly of interest to provide a Natural Language Processing (NLP) pipeline to extract and analyse useful information. Unstructured texts are often noisy, with typing errors and extensive use of non-standard acronyms and medical jargon, which are usually accompanied by a less rigorous structure of the document itself. To overcome these problems, researchers must begin to address issues of spelling correction, acronyms disambiguation and entity normalization. However, in languages other than English it is very difficult to find advanced models, data or other resources.

In this paper we deal with the spelling correction task (i.e. the correction of typos) of notes written by physicians, so as to provide the most correct text for the sophisticated Information Extraction (IE) techniques that generally follow the initial data cleaning phase. Indeed, this work is part of a larger project [7] that involves the Turin's City of Health and Science[2], the Bruno Kessler Foundation[3] and the University of Turin[4]. The project aims at supporting physicians in making decisions in the context of home hospitalization services [8].

Specifically, with this work we introduce a prototype of a spell-checker designed to work on Italian clinical texts. Although it is still a work-in-progress, to the best of our knowledge it represents currently the first and unique study specifically designed to correct medical texts in Italian.

2. Related works

The automatic spelling correction task is the first step to be taken in order to analyse clinical texts, representing one of the most important open problems in Natural Language Processing. The correction process can be divided, according to [9] and [10], in: 1) error detection; 2) correction candidates generation; 3) suggestions ranking. [9] also categorizes errors into two types: non-word errors (errors that give rise to non-existent words in the vocabulary) and real-word errors (when the typo is a meaningful word but not the intended word in that context). In the latter case, particular attention must be paid by developing specific mechanisms, as done for example in [11] and [12].

According to the Claude Shannon's noisy-channel framework, the problem is broken down into *error modelling* (aka *channel model*) and *language modelling*. The first one measure "fitness" of the correction candidate with respect to the corrected string, meanwhile the second one expresses the probability of correct word occurrence, considering - possibly - also the context. Most of the works, such as [13], [14] and [15], use edit distance (Damerau–Levenshtein or Levenshtein distance or Longest common subsequence) for error modelling. Instead, other more refined models make use of word-confusion matrices (calculated from a corpus of typical errors) [16, 17], n-grams of characters [18] or rely on tools such as Aspell[5] which includes phonetic algorithms. In a similar way, it is possible to approach the language model with the

[2] https://www.cittadellasalute.to.it
[3] https://www.fbk.eu/en/
[4] https://ch4i.di.unito.it
[5] http://aspell.net/

simplest n-grams [19], integrating POS tagging [14] or word embeddings [13]. State-of-the-art works [12, 20] still rely on such techniques.

Recently, it has also been shown how good results can be obtained through purely neural approaches. For example, a state-of-the-art corrector for Italian ([21]) uses a biLSTM network for learning the error model and directly correct typos. Unfortunately, in addition to being the only recent work for the Italian language, the errors are artificially generated and therefore they do not fully represent human-like typos. Diametrically opposite is the solution of [22] which uses a Denoising Transformer to learn real error patterns and generate a training set that is as truthful as possible. Unfortunately, solutions of this type require large amounts of data which are difficult to find in languages other than English; the specificities of the medical domain makes the procurement of these resources even more difficult. Indeed, to the best of our knowledge, there are still no public available solutions to correct medical texts in Italian.

3. Proposal

The proposed spell checker prototype is based on Shannon's noisy channel framework described by the equation:

$$\hat{w} = \arg\max_{w \in V} P(w|x) \tag{1}$$

or, by applying Bayes' rule:

$$\hat{w} = \arg\max_{w \in V} P(x|w)P(w) \tag{2}$$

where $\hat{w}$ indicates the best correction for the misspelled word x; word w is selected from a given and finite vocabulary V.

Since the prior probability $P(w)$ carries too little information, we replace it with a Language Model (LM) that involves context $P(w|w_{i-1})$. Finally, we weight the LM with a parameter λ and, for numerically stability reasons, we move on the logarithmic space. The equation is therefore:

$$\hat{w} = \arg\max_{w \in V} logP(x|w) + \lambda logP(w|w_{i-1}) \tag{3}$$

3.1. Data

Unlike English, languages such as Italian are characterized by limited publicly available resources. Furthermore, considering the specificity of clinical language, the availability of medical texts is even more difficult. Medical terms such as surgical procedures, drugs, anatomical parts etc. constitute a very specific vocabulary that differs from what we can normally find in Italian public corpora. It is therefore necessary to find a collection of suitable medical/scientific documents and build new Language Models on them. Following the suggestions of [23] we collected 2M sentences from Wikipedia scientific articles[6], informative articles from the Ministry of Health's website[7], pathologies, drugs and package inserts from Dica33[8] (a popular medical information

[6] https://it.wikipedia.org/ – *Apr 2021 dump*

[7] https://www.issalute.it/index.php/la-salute-dalla-a-alla-z – *retrieved Jun 2021*

[8] https://www.dica33.it/ – *retrieved Jun 2021*

web-site) and – to build a more accurate model of the Italian language – a selection of newspaper articles[9]. Finally, we integrated our corpus with personal medical resumes, which cover most of the subjects studied during the university course. Details on the composition of the corpus are shown in the Table 1. A common feature of the corpora described above is the control over the texts, which limits the presence of typing errors (contrarily to what can happen in a hospital environment).

For computational reasons, and to avoid too rare expressions (a symptom of a possible error) we only consider the elements that occur more than 8 times for the terms and 48 for the n-grams. The resulting vocabulary has a total of 787,940 unique words.

Source	Sentences	Words	
Wikipedia	1,096,672	25,605,524	36%
News	247,872	5,878,905	8%
Ministry of Health	39,838	1,151,371	2%
Dica33	1,059,063	37,333,844	53%
Notes	58,160	962,408	1%
TOTAL	2,501,605	70,932,052	

Table 1
Composition of the constructed corpus.

As regards clinical documents, we relied on a sample of 200 anamnesis notes that were provided by the hospital. The texts, once anonymized, were manually corrected by physicians, thus constituting the gold-standard. Acronyms and abbreviations are excluded from the correction process. Out of the total 9374 words, 269 (2.87%) constitute typos, of which 28 (0.30%) are real-word errors. Errors attributable to purely medical context are 107 (40%), of these 25 are names of drugs/active substances. 88% of typos have Damerau–Levenshtein distance of 1 from the correction (e.g. *mammamria* → *mammaria*); only 10% have DLD 2 (e.g. *ematochici* → *ematochimici*) and a less than 1% for higher distances (e.g. *idrixixizima* → *idroxizina*).

3.2. Model

For simplicity, and considering the rarity with which they occur, we have chosen to discard real-word errors, thus focusing on the remaining 88% of typos. These errors are easily identifiable by searching for terms that do not match the vocabulary. Potential acronyms and abbreviations are excluded; in this regard we have built, in collaboration with domain names, a blacklist of terms not to be corrected.

Once a potential error has been found, a list of candidate corrections is generated, considering the words “similar" to the original one as candidates. Also in this case, the Damerau–Levenshtein distance is used as similarity metric between strings. Since the generation of candidates is a very demanding task in computational terms, we rely on the optimized tools SymSpell[10], which

[9] https://webhose.io/free-datasets/italian-news-articles/ – *Crawled Oct 2015*
[10] https://github.com/wolfgarbe/SymSpell

can operate under the “CLOSEST" regime (i.e. finding the first word at the shortest distance) or “ALL" (all words with maximum distance n).

Always with reference to Equation 3, we list below the solutions that have been implemented and tested.

3.2.1. Error model

The simplest way to implement the channel model is to assign a probability (α) for the event $x = w$ (i.e. the word found is not an error even if it does not appear in the vocabulary) and use the Damerau–Levenshtein distance to evaluate other cases.

$$P(x|w) = \begin{cases} \alpha & \text{if } x = w \\ -log(D(x,w)) & \text{otherwise} \end{cases}$$

In 1991, [24] proposes a simple, but effective, model that uniformly distribute the $1 - \alpha$ probability over all generated candidates $C(x)$. The formula, to which we have added a parameter ϵ as lower-bound, is therefore:

$$P(x|w) = \begin{cases} \alpha & \text{if } x = w \\ \frac{1-\alpha}{|C(x)|} & \text{if } x \in C(x) \\ \epsilon & \text{otherwise} \end{cases}$$

Finally, we tested a slightly more sophisticated variant, proposed by [25], by replacing the uniform distribution with a more informed probability on the characteristics of the language. More specifically, [25] introduced confusion matrices, for each transformation[11], that lists the number of times one character was confused with another one. We can formulate the model as:

$$P(x|w) = \begin{cases} \alpha & \text{if } x = w \\ \Pi_{edit} \frac{edit(x_i,w_j)}{count(x_i,w_j)} & \text{if } x \in C(x) \\ \epsilon & \text{otherwise} \end{cases}$$

3.2.2. Language model

We have chosen to approach the language model in two ways: through n-grams, either monodirectional (w_{i-2}, w_{i-1}, w_i) or bidirectional (w_{i-1}, w_i, w_{i+1}), or through contextualized word embeddings (Masked Language Model). The n-grams are calculated according to the “stupid backoff” scheme [26], having a significant number of tokens available. With regard to the embeddings models, we have experimented with Italian pre-trained BERT-like models such as ELECTRA[27][12], RoBERTa[28][13] and XLMRoberta[29][14] (multi-language RoBERTa that includes Italian).

[11] Transformations allowed by Damerau–Levenshtein are: deletion, insertion, substitution and transposition

[12] dbmdz/electra-base-italian-xxl-cased-generator

[13] idb-ita/gilberto-uncased-from-camembert

[14] xlm-roberta-base

Model	Acc	F1	P	R	TN	FN	FP	TP
Microsoft Office	93.93	21.52	17.26	28.57	8727	195	374	78
Microsoft Office + voc	92.72	33.53	22.84	63.00	8520	101	581	172
Google Docs	98.45	74.96	70.91	79.49	9012	56	89	217
Google Docs + voc	**98.57**	**75.90**	**74.56**	**77.29**	**9029**	**62**	**72**	**211**
LanguageTool	95.71	10.27	13.14	8.42	8949	250	152	23
LanguageTool + voc	96.47	16.20	26.23	11.72	9011	241	90	32
Hunspell	95.31	6.38	7.61	5.49	8919	258	182	15
Hunspell + voc	96.16	9.55	15.20	6.96	8995	254	106	19
N-Grams	**97.98**	**66.31**	**64.58**	**68.13**	**8999**	**87**	**102**	**186**
ELECTRA	97.91	62.31	65.59	59.34	9016	111	85	162
RoBERTa	97.59	58.30	58.74	57.87	8990	115	111	158
XLMRoberta	97.55	57.09	58.17	56.04	8991	120	110	153

Table 2
Results obtained from current models (above) and our proposals (below), where TN: non-error words not corrected; FN: misspelled words not corrected; FP: non-error words erroneously corrected; TP: misspelled words corrected.

4. Results

We tested the different models on the gold standard described above, comparing the results with the state of the art as shown in Table 2. Unfortunately the absence of clinical corpus, as well as publicly available models, makes it difficult, if not impossible, the comparison between spell-checkers in the hospital setting. For this reason we have relied on commonly used general purpose tools such as: Hunspell[15], LanguageTool[16], Google Docs[17] and Microsoft Office[18]. To standardize the results we have chosen to correct all the possible typos suggested by the software, replacing them with the first suggestion. It is also necessary to consider the specificity of the medical vocabulary, which is usually absent in generic tools. For this reason we have excluded all the terms that belong to the vocabulary in our possession and the blacklist defined with the experts. In the Table 2 it is possible to distinguish the two cases (with or without the vocabulary extension) by means of the label "+voc".

As regards the models developed by us, the optimal configuration of the parameters is shown in Table 3. Any change to them does not bring any benefit, as described below. Among the proposed solutions, the best model is the combination of uniform distribution, as error model, and n-grams for the language model.

The advantage of n-grams, over word embeddings, may also be due to the generic nature of the embeddings used; unfortunately, however, the few data available did not allow us to train a

[15] http://hunspell.github.io/
[16] https://languagetool.org/it
[17] https://docs.google.com
[18] https://www.office.com

Parameter	Value
alpha	0.05
lambda	1.0
epsilon	1e-15
n-grams size	5

Table 3
Optimized hyperparameters for the proposed model.

new model from scratch. By excluding Wikipedia and newspapers from the initial corpus, thus focusing on purely medical texts, we can obtain 1 157 061 sentences, useful – after subdivision into training set (96%) and validation set (4%) – to continue the training of the ELECTRA model. However, the effort is vain, failing to differ much from the performance of ELECTRA without fine-tuning; for this reason the results are omitted from the table. The ELECTRA model is however better than other architectures such as RoBERTa and its multilingual variant. A possible advantage in the use of the neural model, in addition to a slight improvement in Precision, consists in the significant reduction of processing times (15 min for the n-grams case vs 25 sec for ELECTRA).

Focusing instead on n-grams, a reduction in their size (passing from 5 to 3) slightly decrease the performances (F1 from 66.31 to 66.19). Similarly, the monodirecional/bidirectional choice is almost irrelevant in terms of score. On the contrary, the application of standardization techniques (stemming and numbers masking) considerably worsens the scores, obtaining F1 62.97. The result is not surprising, as a similar phenomenon has already been observed by [23]. In that case the normalization consisted of lemmatization, but without reliable POS tagging, the lemmatization is de facto reduced to stemming, while the presence of errors, abbreviations and technical terminology makes POS tagging unreliable.

With regard to the error model, on the other hand, the use of the confusion matrix is counterproductive, lowering the F1 score by 8 points on average. While using distance alone as a probability measure has no benefit over the uniform distribution.

5. Discussion and Conclusions

Most of the gaps highlighted in the previous section are probably attributable to the scarcity of data. The examples collected by [23] and [30], for the Italian language, are still too few to be exploited in a machine learning scenario. Meanwhile, synthetic datasets, such as the one used in [21], do not carry with them any useful information to characterize the typical errors of the Italian language. For this reason we are working on the construction of a corpus of common errors for Italian, with the aim of collecting a few thousand pairs $\langle typo, correction \rangle$. Such a dataset would provide the basis for training more sophisticated error models to replace the uniform distribution used in this work.

The texts that characterize the medical domain are also particularly interesting. Just think, for example, that the addition of medical notes (which weighs just over 1%) made it possible to improve performance by 4 points, passing from F1 64.37 to F1 66.31, although with a different

nature of texts. Indeed, all the texts used in our corpus present linguistic characteristics that are very different from those that appear in clinical documents. The Wikipedia entries, for example, provide encyclopedic information, as well as the pages of medical disclosure and personal essays. The search for texts closer to clinical reality will be a fundamental objective that we will pursue in future works.

We also think that - in addition to a mere increase in the size of the datasets for statistical learning purposes - also the integration of syntactic parsing and POS tagging can improve the results. This is especially true in a low resources languages, like Italian, where machine learning is severely limited. For this reason we are conducting a study aimed at evaluating the reliability of these techniques on medical texts.

Finally, the abbreviations (standard and non-standard) usually used in texts remain to be addressed. In this regard it is difficult to think of a pipeline that orders the 3 tasks to be performed: POS tagging, acronyms disambiguation and spelling correction. More likely, the 3 tasks will be carried out in parallel, as one can help the other. We will also evaluate this possibility in future works.

Not having reached the state of the art, represented by Google's spell checker, we believe there is still room for improvement. Moreover, the task is of fundamental importance in order to continue with the analysis of the texts and, ultimately, for clinical decision support.

Acknowledgments

This research has been partially carried out within the "Circular Health for Industry" project funded by "Compagnia di San Paolo" under the call "IA, uomo e società".

References

[1] M. Matarese, M. Lommi, M. G. De Marinis, B. Riegel, A systematic review and integration of concept analyses of self-care and related concepts, Journal of Nursing Scholarship 50 (2018) 296–305.

[2] K. T. Hickey, S. Bakken, M. W. Byrne, D. C. E. Bailey, G. Demiris, S. L. Docherty, S. G. Dorsey, B. J. Guthrie, M. M. Heitkemper, C. S. Jacelon, T. J. Kelechi, S. M. Moore, N. S. Redeker, C. L. Renn, B. Resnick, A. Starkweather, H. Thompson, T. M. Ward, D. J. McCloskey, J. K. Austin, P. A. Grady, Precision health: Advancing symptom and self-management science, Nursing Outlook 67 (2019) 462–475.

[3] F. Alloatti, A. Bosca, L. Di Caro, F. Pieraccini, Diabetes and conversational agents: the aida project case study, Discover Artificial Intelligence 1 (2021).

[4] M. Dumas, W. M. P. van der Aalst, A. H. M. ter Hofstede (Eds.), Process-Aware Information Systems: Bridging People and Software Through Process Technology, Wiley, 2005.

[5] E. Rojas, J. Munoz-Gama, M. Sepúlveda, D. Capurro, Process mining in healthcare: A literature review, Journal of biomedical informatics 61 (2016) 224–236.

[6] I. A. Amantea, E. Sulis, G. Boella, R. Marinello, D. Bianca, E. Brunetti, M. Bo, C. Fernández-Llatas, A process mining application for the analysis of hospital-at-home admissions, in: L. B. Pape-Haugaard, C. Lovis, I. C. Madsen, P. Weber, P. H. Nielsen, P. Scott (Eds.), Digital

Personalized Health and Medicine - Proceedings of MIE 2020, Medical Informatics Europe, Geneva, Switzerland, April 28 - May 1, 2020, volume 270 of *Studies in Health Technology and Informatics*, IOS Press, 2020, pp. 522–526.

[7] R. Aringhieri, G. Boella, E. Brunetti, L. D. Caro, C. D. Francescomarino, M. Dragoni, R. Ferrod, C. Ghidini, R. Marinello, M. Ronzani, E. Sulis, Towards the application of process mining for supporting the home hospitalization service, in: A. Marrella, D. T. Dupré (Eds.), Proceedings of the 1st Italian Forum on Business Process Management co-located with the 19th International Conference of Business Process Management (BPM 2021), Rome, Italy, September 10th, 2021, volume 2952 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 33–38.

[8] I. A. Amantea, M. Arnone, A. D. Leva, E. Sulis, D. Bianca, E. Brunetti, R. Marinello, Modeling and simulation of the hospital-at-home service admission process, in: M. S. Obaidat, T. I. Ören, H. Szczerbicka (Eds.), Proceedings of the 9th International Conference on Simulation and Modeling Methodologies, Technologies and Applications, SIMULTECH 2019, Prague, Czech Republic, July 29-31, 2019, SciTePress, 2019, pp. 293–300.

[9] K. Kukich, Techniques for automatically correcting words in text, ACM Comput. Surv. 24 (1992) 377–439.

[10] T. A. Pirinen, K. Lindén, State-of-the-art in weighted finite-state spell-checking, in: CICLing, 2014.

[11] S. Deorowicz, M. Ciura, Correcting spelling errors by modelling their causes, International Journal of Applied Mathematics and Computer Science 15 (2005) 275–285.

[12] C. Whitelaw, B. Hutchinson, G. Chung, G. Ellis, Using the web for language independent spellchecking and autocorrection, in: EMNLP, 2009.

[13] G. Damnati, J. Auguste, A. Nasr, D. Charlet, J. Heinecke, F. Bechet, Handling Normalization Issues for Part-of-Speech Tagging of Online Conversational Text, in: Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 2018.

[14] J. Dziadek, A. Henriksson, M. Duneld, Improving terminology mapping in clinical text with context-sensitive spelling correction, Studies in Health Technology and Informatics 235 (2017) 241–245.

[15] A. Sorokin, T. Shavrina, Automatic spelling correction for russian social media texts, 2016.

[16] Y. Lv, Y. Deng, M. Liu, Q. Lu, Automatic error checking and correction of electronic medical records, in: G. Chen, F. Liu, M. Shojafar (Eds.), Fuzzy System and Data Mining - Proceedings of FSDM 2015 [Shanghai, China, December 12-15, 2015], volume 281 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2015, pp. 32–40.

[17] M. Banko, E. Brill, Scaling to very very large corpora for natural language disambiguation, in: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Toulouse, France, 2001, pp. 26–33.

[18] J. Vilares, M. A. Alonso, Y. Doval, M. Vilares, Studying the effect and treatment of misspelled queries in cross-language information retrieval, Inf. Process. Manage. 52 (2016) 646–657.

[19] G. Héja, G. Surján, Using n-gram method in the decomposition of compound medical diagnoses, International journal of medical informatics 70 (2003) 229—236.

[20] J. Gupta, Z. Qin, M. Bendersky, D. Metzler, Personalized online spell correction for personal search, in: The World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2785–2791.

[21] L. Sbattella, R. Tedesco, How to simplify human-machine interaction: A text complexity calculator and a smart spelling corrector, in: Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good, Goodtechs '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 304–305.

[22] A. Kuznetsov, H. Urdiales, Spelling correction with denoising transformer, 2021. arXiv:2105.05977.

[23] E. Mensa, G. M. Marino, D. Colla, M. Delsanto, D. P. Radicioni, A resource for detecting misspellings and denoising medical text data, in: J. Monti, F. dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020.

[24] E. Mays, F. J. Damerau, R. L. Mercer, Context based spelling correction, Information Processing & Management 27 (1991) 517–522.

[25] M. D. Kernighan, K. W. Church, W. A. Gale, A spelling correction program based on a noisy channel model, in: COLING, 1990.

[26] T. Brants, A. C. Popat, P. Xu, F. J. Och, J. Dean, Large language models in machine translation, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 858–867.

[27] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: ICLR, 2020.

[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[29] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. arXiv:1911.02116.

[30] M. Hagiwara, M. Mita, Github typo corpus: A large-scale multilingual dataset of misspellings and grammatical errors, in: LREC, 2020.