A Resource for Detecting Misspellings and Denoising Medical Text Data

Enrico Mensa[⋆]
Università di Torino,
Dipartimento di Informatica

Gian Manuel Marino[†]
Università di Torino,
Dipartimento di Informatica

Davide Colla[⋆]
Università di Torino,
Dipartimento di Informatica

Matteo Delsanto[⋆]
Università di Torino,
Dipartimento di Informatica

Daniele P. Radicioni[⋆]
Università di Torino,
Dipartimento di Informatica

[⋆]{firstname.surname}@unito.it; [†]marino.jnf@gmail.com

Abstract

English. In this paper we propose a method for collecting a dictionary to deal with noisy medical text documents. The quality of such Italian Emergency Room Reports is so poor that in most cases these can be hardly automatically elaborated; this also holds for other languages (e.g., English), with the notable difference that no Italian dictionary has been proposed to deal with this jargon. In this work we introduce and evaluate a resource designed to fill this gap.[1]

Italiano. *In questo lavoro illustriamo un metodo per la costruzione di un dizionario dedicato all'elaborazione di documenti medici, la porzione delle cartelle cliniche annotata nei reparti di pronto soccorso. Questo tipo di documenti è così rumoroso che in genere le cartelle cliniche difficilmente posono essere direttamente elaborate in maniera automatica. Pur essendo il problema di ripulire questo tipo di documenti un problema rilevante e diffuso, non esisteva un dizionario completo per trattare questo linguaggio settoriale. In questo lavoro proponiamo e valutiamo una risorsa finalizzata a condurre questo tipo di elaborazione sulle cartelle cliniche.*

1 Introduction

Noise in textual data is a very common phenomenon afflicting text documents, especially when dealing with informal texts such as chats, SMS and e-mails. This kind of text inherently contains spelling errors, special characters, non-standard word forms, grammar mistakes, and so on (Liu et al., 2012). In this work we focus on a type of text which can also be very noisy: *emergency room reports*. In the broader frame of a project aimed at detecting injuries stemming from violence acts in narrative texts contained in emergency room reports, we recently developed the VIDES, so dubbed after 'Violence Detection System' (Mensa et al., 2020). This system is concerned with categorizing textual descriptions as containing violence-related injuries (V) vs. non-violence-related injuries (NV), which is a relevant task to the ends of devising alerting mechanisms to track and prevent violence episodes. VIDES combines a neural architecture which performs the categorization step (thus discriminating V and NV records) and a Framenet-based approach, whereby semantic roles are represented through a synthetic description employing a set of word embeddings.[2] More specifically, a model of violent event has been devised: records that are recognized as containing violence-related injuries are further processed by an explanation module, which is charged to individuate the main elements corroborating that categorization (V) by identifying the involved agent, the type of injury, the involved body district *etc.*. Explaining the categorization ultimately involves filling the semantic components of the violence frame. All such ele-

[2]Related approaches have been designed as Semantic Role Labeling tasks (Gildea and Jurafsky, 2002; Zapirain et al., 2013), but also frame-based approaches have been proposed, paired to deep syntactic analysis, to extract salient information through a template-filling approach (Lesmo et al., 2009; Gianfelice et al., 2013).

ments contribute to recognizing a violent event as the source of the injuries complained by ER patients.

During the development of VIDES we realized that in order to run sophisticated algorithms for the detection and extraction of such violent traits we needed to cope with the noise contained in the input medical records. Some efforts have been invested to deal with different sorts of linguistic phenomena menacing the comprehension of texts; however, most existing works are focused on the English language, and rely on dictionaries that cannot be directly employed on Italian text documents.

In this preliminary work we start to tackle the issue of noisy words in medical records for Italian texts, by specifically focusing on misspellings. Our contribution is twofold: we first manually explore the dataset by analyzing a small sample of records in order to determine whether the main traits and issues present in other languages are also shared by Italian reports; secondly, we collect, merge and evaluate a set of Italian dictionaries, which constitute a brick fundamental to build any domain specific spell-checking algorithm (López-Hernández et al., 2019).

2 Related Work

Literature shows a limited but significant interest on the issue of detecting and correcting noisy medical text documents; nonetheless, some commonalities underlying this sort of text can be drawn.

Medical texts are often very noisy; among the most common mistakes we mention mistyping, lack or improper use of punctuation, grammatical errors and domain-specific abbreviations and Latin medical terminology (Siklósi et al., 2013). This is mainly due to the nature of the records themselves, and to the fact that the medical personnel compiling the entries is often under pressure and in a hurry.

Most of the spelling correction approaches have been carried out for English, with the exception of research in Swedish (Dziadek et al., 2017) and Hungarian (Siklósi et al., 2013), while no work has been found dealing with the Italian language. Regarding the methodologies, most works focus on non-word errors, while disregarding grammatical and real word mistakes. Non-word mistakes occur when a misspelling error produces a word that does not exist, such as 'patienz' instead of 'patient', while real word mistakes occur when a word is mistakenly replaced with another –existing– one, like the substitution of 'abuse' with 'amuse'. The adopted algorithms are diverse, with the prevalence of approaches relying on embeddings (Kilicoglu et al., 2015; Workman et al., 2019) or regular expressions and rule-based systems (Patrick et al., 2010; Sayle et al., 2012; Lai et al., 2015). However, basically all contributions adopt a preliminary *dictionary look-up* step (López-Hernández et al., 2019). To this purpose, besides the general dictionaries provided in toolkits such as Aspell and Google Spell Checker,[3] authors often rely on (medical) domain-specific dictionaries, such as The Unified Medical Language System (UMLS) (Aoki et al., 2004), the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT, 2020) and The SPECIALIST Lexicon (Browne et al., 2000). It is thus evident that the development of analogous resources for the Italian language is a crucial step for the design of tools and systems aimed at dealing with the spell-checking of Italian medical text documents.

Besides the treatment of misspellings, there are also works specifically focused on abbreviations. For instance, in (Wu et al., 2011) the authors present a corpus-based method to create a lexical resource of English clinical abbreviations via several machine learning algorithms. The resource has been used to automatically detect and expand abbreviations, and obtained interesting experimental results. More recently, another approach proposed in (Kreuzthaler et al., 2016) focuses on abbreviations ending with a period character; the proposed technique puts together statistical and dictionary-based strategies to detect abbreviations in German clinical narratives.

In the present work we are not proposing a specific technique for dealing with abbreviations, we are rather concerned with misspellings. However, the approaches already proposed for other languages will be considered in future work to also treat Italian abbreviations in our dataset.

3 Data Analysis

We analyze real data coming from a set of emergency room reports collected in Italian hospitals by the Italian National Institute of Health in the

[3] `https://git.savannah.gnu.org/git/aspell.git` and `https://languagetool.org`, respectively.

Table 1: Figures describing the complete dataset and the sample selected for manual annotation.

	Complete	Sample
Number of entries	136,144	592
Number of tokens	2,329,840	14,137
Number of unique tokens	49,116	1,842
Avg tokens per record	17.11	23.88

frame of the SINIACA project (Pitidis et al., 2014). The SINIACA project, so dubbed after 'Sistema Informativo Nazionale sugli Incidenti in Ambiente di Civile Abitazione' (National Information System on Accidents in Civil Housing Environment), is the Italian branch of the European Injury Database (EU-IDB) (Lyons et al., 2015), an EU-wide surveillance system concerned with accidents, collecting data from hospital emergency departments according to the EU recommendation no. C 164/2007/01, aimed at injury prevention and safety promotion.

Dataset. The whole dataset amounts to 136,144 non-empty entries, 592 of which were randomly selected for the manual analysis. Table 1 reports some figures describing the dataset. Double spaces and punctuation redundancy have been fixed through regular expressions, while tokens have been extracted by splitting the sentences based on spaces. Also, tokens containing numbers are presently discarded.

Analysis result. We performed a manual analysis on the subset of the original dataset: the 592 randomly selected entries herein were manually examined, and for each entry we looked for noisy words. Three main types of words were annotated: *i*) **misspellings**: a wrongly typed word, e.g., *fratura* instead of *frattura* – fracture; *ii*) **abbreviations**: a shortened form of a word or phrase, e.g., *dx* instead of *destra* – right; *iii*) **acronyms**: a word formed from the initial letters of other words, e.g., *ps* instead of *pronto soccorso* – emergency room. Interestingly enough, both abbreviations and acronyms can be at least partly considered as domain dependent: for example, in different settings, *ps* may denote *post scriptum* (something added at a later time, likely a letter, after the signature), but also 'Polizia di Stato' (Police) or 'previdenza sociale' (social security). Dealing with such phenomena thus involves accessing a context dependent knowledge base that allows selecting the utterance appropriate for the context at hand. We are presently concerned with misspellings, acronyms and abbreviations as *noise*, but only the first category can be actually considered as an error. More specifically, while misspellings are actual errors, abbreviations and acronyms belong to a domain-specific language, and these are way too specific to be recognized as legitimate words through a general-purpose dictionary. As seen in literature, misspellings and abbreviations/acronyms must be treated with different techniques, and in this work we mainly focus on tackling the first category, while also obtaining interesting insights regarding the second one.

Table 2: Noise distribution on the annotated dataset; between parenthesis we report the percentage over the total number of tokens, while the last column indicates the average per record.

	With repetitions	Unique	Average
Noisy tokens	1,336 (9.4%)	424 (3%)	2.25
Misspellings	433 (3%)	304 (2%)	0.73
Abbreviations	670 (4.7%)	76 (0.5%)	1.13
Acronyms	233 (1.6%)	45 (0.3%)	0.39

Table 2 illustrates the results of the annotation process. We discovered that the dataset contains a lot of noise, amounting to almost the 10% of the tokens, on average 2 noisy tokens per record. By looking separately at the different typologies of noise we observe that misspellings are more scattered and diverse, while the usage of abbreviations and acronyms seems to be more coherent: we have 670 instances of abbreviations but only 76 unique abbreviations, while 304 out of the 433 instances of misspellings are unique. This phenomenon is also depicted in Figure 1, where we provide the log-log plot of the frequency of each misspelling, abbreviation and acronym ordered by rank. We observe that the distribution of abbreviations and acronyms has a different magnitude, but is very similar in shape; on the other side, the misspellings are clearly more scattered with a very long tail of items appearing only once.

4 Dictionaries Creation and Evaluation

The manual analysis uncovered characteristics and features that are in line with those found in literature for English datasets (López-Hernández et al., 2019). However, to allow the development

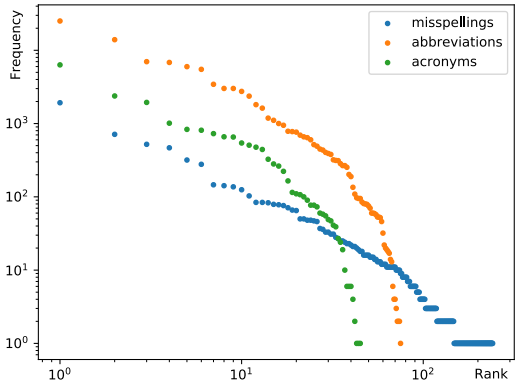

Figure 1: Log-log plot showing the frequency of misspellings, abbreviations and acronyms over the annotated dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of spell-checkers for Italian medical texts, another key component is still missing: most approaches aimed at error detection rely on dictionaries to determine if a token is a legitimate word or not. In fact, the simplest implementation of misspellings detection is as follows: if we have at our disposal the set W containing all of the terms of a given language, joined to all terms pertaining the specific domain at hand, any word $w \notin W$ can be likely considered as a misspell. To the best of our knowledge, no such dictionary exists that is able to cope with Italian medical text documents, so we built a resource to answer to this need.

4.1 Source Dictionaries

The automatic development of a dictionary is not a trivial task. We want to reach the highest possible coverage for both general terms and specific medical terminology, but at the same time we cannot rely too much on unverified sources (e.g., crowd-sourced data) with the risk of introducing misspellings and errors into the dictionary. We selected different sources and arranged them into four main classes:

- `MED`: a collection of medical terms built by putting together five medical online dictionaries (torrinomedica.it, 2020; abcsalute.it, 2020; codifa.it, 2020; my-personaltrainer.it, 2020a; my-personaltrainer.it, 2020b), containing medical specific terms and medications names;

- `ITA`: a collection of Italian terms built by merging three well-known Italian online dictionaries (Hoepli, 2020; Sabatini-Colletti, 2020; De Mauro, 2020);

- `WMED`: a collection of terms from Wikipedia pages pertaining the medical domain. The list of Wikipedia medical pages has been obtained by querying the SPARQL endpoint of Wikidata (Vrandečić and Krötzsch, 2014), while the pages have been taken from the 20 August 2020 Wikipedia dump;

- `WMOV`: since medical records also contain a brief narrative text of the events that led to the (either violent or accidental) injuries, we added terms associated to eventive and narrative genres by collecting Wikipedia pages pertaining to movies, television series and literary work that are expected to contain narrative terminology.

Table 3: Figures of the 5000 annotated tokens used to evaluate the dictionaries.

Class	Type	Amount (% on total)
Positive	Correct words	3,886 (77.7%)
	Abbreviations	184 (3.7%)
	Acronyms	126 (2.5%)
Negative	Misspells	804 (16.1%)

The set of terms extracted from Wikipedia can potentially contain misspellings and errors, and so we also set a frequency minimum which allows for the pruning of the tokens herein. We annotate this parameter with a subscript next to the set name, e.g., $\texttt{WMOV}_1$ indicates that the threshold was set to 1 for the terms frequency.

4.2 Evaluation

Building the dataset. In order to assess the quality of the collected dictionaries we started from the 49,116 unique tokens in the dataset, removed the stop words[4] and randomly selected 5,000 of them to be manually annotated. The annotation was carried out by four of the authors of this paper. The selection algorithm was designed so to increase the probability of a token to be selected in accordance to its frequency in the dataset. These 5,000 tokens were then annotated

[4]We used the set of stop words made available by Spacy (`https://spacy.io/`) for the Italian language.

Table 4: Results of the evaluation of the considered dictionaries. The first column reports the size of each dictionary, the second to fourth columns provide coverage and correctness along with their harmonic mean, while the last three columns illustrate the coverage of our dictionaries on tokens that were annotated as correct words, abbreviations and acronyms.

	Terms	Coverage	Correctness	F1-Score	Correct Words	Abbreviations	Acronyms
ITA	124,494	.542	**.980**	.700	.573	.179	.206
MED, ITA	155,650	.621	.975	.759	.652	.228	.261
MED, ITA, WMED_0	287,279	.897	.907	.902	.918	.521	.785
MED, ITA, WMED_0, WMOV_0	511,827	**.926**	.863	.894	.941	.641	.873
MED, ITA, WMED_1, WMOV_1	343,264	.906	.898	.902	.925	.586	.793
MED, ITA, WMED_1, WMOV_5	266,633	.892	.922	**.907**	.912	.554	.761
(*LEM*) MED, ITA, WMED_1, WMOV_5	227,895	.903	.896	.900	.926	.559	.674

with one of the following four classes: correct words (regardless of their domain specificity), abbreviations, acronyms and misspellings. The first three classes represent terms that should be found in our resource, while the last category contains words that should not be present in the dictionary. Table 3 reports the statistics featuring the dataset annotated for evaluation purposes.

Evaluating the dictionary. In Table 4 we report the results of the dictionaries evaluation. Each dictionary has been built by taking into consideration one or more of the previously presented sources. Multiple sources have been simply merged into a unique set of terms, without repetitions. We assess the quality of each dictionary via two measures, coverage and correctness. The *coverage* is the percentage of words that were found in the dictionary (either correct words, abbreviations or acronyms), while the *correctness* is the percentage of misspellings that were not present in the dictionary. We considered different combinations of the sources, the tuning of the frequency-based filtering parameter, and an additional lemmatization step.

We observe that both the ITA and the MED sets are fundamentally correct, even though they also include words that in the common usage are frequently misspelled, such as *passeggiero* in place of the correct form *passeggero*. On the other side, its .62 coverage is unsatisfactory (please refer to the second row of Table 4, MED, ITA); it also witnesses that medical jargon is only partially grasped by dictionaries in the MED set. As expected, the introduction of terms from Wikipedia improves the coverage, but with detrimental effect on the correctness. This also holds for the WMOV set, which is rich but also pretty noisy. By fine tuning the frequency thresholds of both WMED and WMOV we were able to prune most of the noise and to preserve the coverage at the same time, finally obtaining a good dictionary with the combination MED, ITA, WMED_1, WMOV_5.

This setting was also tested by applying a lemmatization step on both Wikipedia terms and our dataset tokens. Interestingly, the lemmatization introduces more mistakes than it solves: this is due the the fact that unpredictably the lemmatizer converts misspellings into legitimate words that do not necessarily correspond to their correct spelling. This fact shows also that lemmatization, which is acknowledged as a task almost completely solved from a scientific point of view, still poses relevant issues for the medical jargon and for domain-specific languages more in general.

A lot of abbreviations are not yet covered in the dictionary. Once again, these abbreviations are dataset-specific (and perhaps also follow local uses rather than widely accepted practices), and thus these are very hard to find even on specialized public medical resources. For instance, *incid* (*incidente* – accident) appears very frequently and its easily understandable by humans but its not a common or medical abbreviation. The same phenomenon can also be observed on acronyms, that are less sparse and more adherent to widely accepted practices and standards.

5 Conclusions and Future Work

In this work we tackled the issue of detecting textual noise in Italian room emergency reports, focusing specifically on misspellings. Firstly we examined the reports and found out that the sorts of issues reported in literature for other languages can also be found in Italian text documents. Secondly, we developed and evaluated an Italian dictionary suited for the task of noise detection. In future work we plan to expand the dictionary by

including the terms from the Italian ICD-9 and ICD-10 (International Classification of Diseases), that may be useful to interpret acronyms and resolve abbreviations. Moreover, we plan to employ this dictionary in a fully fledged spell-checking system. Finally, the usage of semantic —sense indexed— representations such as, e.g., (Mensa et al., 2018) and (Colla et al., 2020a; Colla et al., 2020b) will be explored, in order to deal with real word mistakes, and more in general contextual information (Basile et al., 2019) will be considered as a main cue in order to uncover and correct this sort of errors. For example, by leveraging the terminology surrounding noisy tokens we plan to distinguish the more scattered misspellings from the other terms that are not present in our dictionary.

Acknowledgments

The first author was supported by a grant provided by Università degli Studi di Torino. This research is also supported by Fondazione CRT, RF 2019.2263.

References

[abcsalute.it2020] abcsalute.it. 2020. Abcsalute.it - Dizionario Medico. http://www.abcsalute.it/dizionario-medico.

[Aoki et al.2004] Kiyoko F Aoki, Atsuko Yamaguchi, Nobuhisa Ueda, Tatsuya Akutsu, Hiroshi Mamitsuka, Susumu Goto, and Minoru Kanehisa. 2004. Kcam (kegg carbohydrate matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic acids research*, 32(suppl_2):W267–W272.

[Basile et al.2019] Valerio Basile, Tommaso Caselli, and Daniele P. Radicioni. 2019. Meaning in Context: Ontologically and linguistically motivated representations of objects and events. *Applied Ontology*, 14:335–341.

[Browne et al.2000] Allen C Browne, Alexa T McCray, and Suresh Srinivasan. 2000. The specialist lexicon. *National Library of Medicine Technical Reports*, pages 18–21.

[codifa.it2020] codifa.it. 2020. codifa.it - Dizionario dei Farmaci. https://www.codifa.it/farmaci.

[Colla et al.2020a] Davide Colla, Enrico Mensa, and Daniele P. Radicioni. 2020a. Lesslex: Linking multilingual embeddings to sense representations of lexical items. *Computational Linguistics*, 46(2):289–333.

[Colla et al.2020b] Davide Colla, Enrico Mensa, and Daniele P. Radicioni. 2020b. Novel metrics for computing semantic similarity with sense embeddings. *Knowledge-Based Systems*, 206:106346.

[De Mauro2020] De Mauro. 2020. Dizionario Italiano Nuovo De Mauro. https://dizionario.internazionale.it/.

[Dziadek et al.2017] Juliusz Dziadek, Aron Henriksson, and Martin Duneld. 2017. Improving terminology mapping in clinical text with context-sensitive spelling correction. *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, 235:241.

[Gianfelice et al.2013] Davide Gianfelice, Leonardo Lesmo, Monica Palmirani, Daniele Perlo, and Daniele P. Radicioni. 2013. Modificatory Provisions Detection: a Hybrid NLP Approach. In Bart Verheij, editor, *Proceedings of ICAIL 2013: XIV International Conference on Artificial Intelligence and Law*, pages 43–52. ACM.

[Gildea and Jurafsky2002] Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

[Hoepli2020] Hoepli. 2020. Dizionario Italiano Hoepli. https://dizionari.repubblica.it/italiano.html.

[Kilicoglu et al.2015] Halil Kilicoglu, Marcelo Fiszman, Kirk Roberts, and Dina Demner-Fushman. 2015. An ensemble method for spelling correction in consumer health questions. In *AMIA Annual Symposium Proceedings*, volume 2015, page 727. American Medical Informatics Association.

[Kreuzthaler et al.2016] Markus Kreuzthaler, Michel Oleynik, Alexander Avian, and Stefan Schulz. 2016. Unsupervised abbreviation detection in clinical narratives. In *Proceedings of the clinical natural language processing workshop (ClinicalNLP)*, pages 91–98.

[Lai et al.2015] Kenneth H Lai, Maxim Topaz, Foster R Goss, and Li Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. *Journal of biomedical informatics*, 55:188–195.

[Lesmo et al.2009] Leonardo Lesmo, Alessandro Mazzei, and Daniele P. Radicioni. 2009. Extracting Semantic Annotations from Legal Texts. In *Proceedings of the International Conference on Hypertext, HT09*, pages 167–172, Turin, Italy, July. ACM.

[Liu et al.2012] Fei Liu, Fuliang Weng, and Xiao Jiang. 2012. A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1035–1044.

[López-Hernández et al.2019] Jésica López-Hernández, Ángela Almela, and Rafael Valencia-García. 2019. Automatic spelling detection and correction in the medical domain: A systematic literature review. In *International Conference on Technologies and Innovation*, pages 95–108. Springer.

[Lyons et al.2015] Ronan Lyons, Rupert Kisse, and Wim Rogmans. 2015. Eu-injury database introduction to the functioning of the injury database (idb). https://bit.ly/37FAKaB.

[Mensa et al.2018] Enrico Mensa, Daniele P. Radicioni, and Antonio Lieto. 2018. Cover: a linguistic resource combining common sense and lexicographic information. *Language Resources and Evaluation*, 52(4):921–948.

[Mensa et al.2020] Enrico Mensa, Davide Colla, Marco Dalmasso, Marco Giustini, Carlo Mamo, Alessio Pitidis, and Daniele P. Radicioni. 2020. Violence detection explanation via semantic roles embeddings. *BMC Medical Informatics and Decision Making*, 20(1):263–275, Oct.

[my-personaltrainer.it2020a] my-personaltrainer.it. 2020a. Lista delle Malattie di My Personal Trainer. https://www.my-personaltrainer.it/malattie_a_z.php.

[my-personaltrainer.it2020b] my-personaltrainer.it. 2020b. Lista di Sintomi di My Personal Trainer. https://www.my-personaltrainer.it/sintomi_a_z.php.

[Patrick et al.2010] Jon Patrick, Mojtaba Sabbagh, Suvir Jain, and Haifeng Zheng. 2010. Spelling correction in clinical notes with emphasis on first suggestion accuracy. In *Proceedings of 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2010)*, pages 1–8.

[Pitidis et al.2014] Alessio Pitidis, Gianni Fondi, Marco Giustini, Eloïse Longo, Giuseppe Balducci, Gruppo di lavoro SINIACA-IDB, and Dipartimento di Ambiente e Connessa Prevenzione Primaria, ISS. 2014. Il Sistema SINIACA-IDB per la sorveglianza degli incidenti. *Notiziario dell'Istituto Superiore di Sanità*, 27(2):11–16.

[Sabatini-Colletti2020] Sabatini-Colletti. 2020. Dizionario Italiano Sabatibi Colletti. https://dizionari.corriere.it/dizionario_italiano/.

[Sayle et al.2012] Roger Sayle, Paul Hongxing Xie, and Sorel Muresan. 2012. Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction. *Journal of chemical information and modeling*, 52(1):51–62.

[Siklósi et al.2013] Borbála Siklósi, Attila Novák, and Gábor Prószéky. 2013. Context-aware correction of spelling errors in hungarian medical documents. In *International Conference on Statistical Language and Speech Processing*, pages 248–259. Springer.

[SNOMED-CT2020] SNOMED-CT. 2020. International Health Terminology Standards Development Organisation. http://www.ihtsdo.org/snomed-ct/.

[torrinomedica.it2020] torrinomedica.it. 2020. torrinomedica.it - Dizionario dei Farmaci. https://www.torrinomedica.it/schede-farmaci.

[Vrandečić and Krötzsch2014] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September.

[Workman et al.2019] T Elizabeth Workman, Yijun Shao, Guy Divita, and Qing Zeng-Treitler. 2019. An efficient prototype method to identify and correct misspellings in clinical text. *BMC research notes*, 12(1):1–5.

[Wu et al.2011] Yonghui Wu, S Trent Rosenbloom, Joshua C Denny, Randolph A Miller, Subramani Mani, Dario A Giuse, and Hua Xu. 2011. Detecting abbreviations in discharge summaries using machine learning methods. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1541. American Medical Informatics Association.

[Zapirain et al.2013] Benat Zapirain, Eneko Agirre, Lluis Marquez, and Mihai Surdeanu. 2013. Selectional preferences for semantic role classification. *Computational Linguistics*, 39(3):631–663.