Automatic Extraction of Polish Language Errors from Text Edition History

Roman Grundkiewicz

Adam Mickiewicz University,
Faculty of Mathematics and Computer Science,
ul. Umultowska 87, 61-614 Poznan, Poland
romang@amu.edu.pl

Abstract. There are no large error corpora for a number of languages, despite the fact that they have multiple applications in natural language processing. The main reason underlying this situation is a high cost of manual corpora creation. In this paper we present the methods of automatic extraction of various kinds of errors such as spelling, typographical, grammatical, syntactic, semantic, and stylistic ones from text edition histories. By applying of these methods to the Wikipedia's article revision history, we created the large and publicly available corpus of naturally-occurring language errors for Polish, called PlEWi. Finally, we analyse and evaluate the detected error categories in our corpus.

Keywords: error corpora, language errors detection, mining Wikipedia

1 Introduction

Error corpora are widely applied in the natural language processing, especially in the course of developing proofreading tools. Gathering the corpus containing annotated naturally-occurring errors in the traditional way is very costly, because it usually entails the manual annotation of text. Consequently, there are no large digital error corpora for a number of languages, as is the case with the Polish language. Admittedly, there exist error corpora of foreign language learners (mainly of English language learners), but non-native errors are quite different [6] and tools developed based on such data may not be sufficiently robust to detect errors made by native-speakers.

To reduce the time and cost of manual work required for collecting language mistakes, corrections made by teachers in written assignments or the history of text editions are subject to analysis. The acquisition of such documents, especially in the electronic form, poses a major challenge as edition history is usually not stored. The exceptions are Wikipedia and other Wiki family members (e.g. WikiNews and other smaller wiki-like websites), services such as Google Docs or even files inside the control version systems.

In this paper we will present the automatic method used for building the corpus of naturally-occurring language errors from the Polish Wikipedia, called PlEWi (*Polish Language Errors from Wikipedia*). In Sect. 2 we will describe the technical aspects of Wikipedia mining and we will present our solution for the detection and extraction of language errors from edition history. Finally, we will analyse and evaluate the collected data in Sections 3 and 4.

1.1 Wikipedia as source of language errors

The advantages of Wikipedia are its size, availability and the contribution made by a diversified community. Moreover, its content is generally considered reliable [9]. But as has been pointed out by Miłkowski [8], Wikipedia probably cannot represent the average language due to its digital form, rather formal and restricted style, uncommon scope of topics and a higher education level of its users. The mere encyclopaedic style of Wikipedia's can be viewed as inconvenient, because some changes are imposed only by the style unification requirements.

Nevertheless, Wikipedia can be perceived as an accurate source of language error corrections as the aim of the majority of its editors is to improve the quality of the content of articles[1]. A high average education level of Wikipedia users may confirm this statement. Please note, that Wikipedia with its community pages is also an up-to-date record of the living language.

1.2 Related works

The idea of using edition histories of documents for the purpose of collecting certain types of language errors abounds in literature. Miłkowski [8] proposed the building of error corpora using Wikipedia revisions based on the hypothesis that the majority of frequent minor edits are the corrections of spelling, grammar, style and usage mistakes. This hypothesis, although very accurate, does not yield the expected result in the form of a wide range of error types, e.g. inflectional errors, because they are rarely repeated.

The work of Max and Wisniewki's [7] has led to the creation of WiCoPaCo — a corpus of naturally-occurring corrections and paraphrases[2]. One of the applications of the WiCoPaCo was the construction of a set of spelling error corrections and its application in the evaluation of the spell checker. However, it did not contain certain types of errors, such as repetitions or omissions of words, and corrections that refers only to punctuation or case modification.

Zesch [10] extracted the samples of real-word spelling errors and their contexts from Wikipedia's revision histories. Collected data were used to evaluate statistical and knowledge-based measures applied in contextual fitness in the task of real-word spell checking. He confirms the opinion that such natural errors are better suited for evaluation purposes than artificially created ones.

2 Extracting language errors

We have accessed the Wikipedia data with script iterates over each two adjacent revision in every article on Wikipedia's dump file in XML format[3]. Edited text fragments from these revisions were extracted using the longest common subsequence (LCS) algorithm and cleaned from the Wikipedia format markups. Next PSI-Toolkit toolbox [3] has been used on all fragments for sentence segmentation and lemmatization in the further stage

[1] The issue of vandalism will be discussed in Sect. 2.3.

[2] `http://wicopaco.limsi.fr/`

[3] `http://dumps.wikimedia.org/plwiki/`

as well. All editions involving only the addition or deletion of the article content were disregarded.

After that, if two edited sentences met certain surface conditions, such as (1) the sentence length is between 4 and 80 tokens, (2) the difference in length is less than 4 tokens, (3) a ratio of words to non-word tokens is higher than 0.75, and (4) a number of non-letter characters is less than a quarter of all the characters, the LCS algorithm was run again. For the time being it worked on tokens instead of lines, so we obtained all edition instances per each sentence.

2.1 Language errors recognition

As a result of the initial stage, for each sentence pair it is obtained the sequence of editions $((u_0, v_0), (u_1, v_1), \ldots)$, where each edition (u, v) is a pair of the older and the newer word(s).

Because too many editions in a sentence imply a rewording or an extension of the sentence rather than error corrections, we rejected the sentences containing more than four editions. But there is no restriction to the mere single word and non-empty editions (i.e. u, v may consist of two words or be empty)[4]. Next, each edition is classified into a defined error category (modeled on Bušta's work [1]) through hand-crafted heuristics, or rejected.

Simple errors First, the word u and its edition v are tested using surface conditions that do not imply the use of any natural language processing tool. In particular, the edition can be easily discarded if (1) v occurs in the list of vulgarisms, (2) u and v differ only in more than one punctuation mark (e.g. *what*→*what???*), (3) the change involves abnormal case modifications like *word*→*WoRd*.

The following types of errors are detected: (1) misused punctuation marks (e.g. missing of a comma or a full stop), (2) misspellings connected with separable and inseparable writing when u and v differ only in the space character or the hyphen, (3) wrong letter case (e.g. *polska*→*Polska* [*Poland*]).

Spelling errors If the edition has not been recognised as error correction by surface conditions, it is labeled using the spell checker[5] with a dictionary D as (1) a non-word spelling correction if $u \notin D$ and $v \in D$, as (2) a real-word error correction if $u, v \in D$, as (3) an act of vandalism if $u \in D$ and $v \notin D$, and as (4) "out of dictionary" if $u, v \notin D$.

For the non-word spelling corrections, there is made a distinction between the misspellings involving in the omission of diacritical signs and the other misspellings. The real-word editions are further classified into one of the grammatical error types, whereas the editions appearing to be examples of vandalism are discarded. As Kukich's studies showed [5], most of language errors are in the short edit distance. Hence, in the case

[4] Further in this work we will use a term *word* even if it is a sequence of words.

[5] We used Hunspell spell checker: `http://hunspell.sourceforge.net/`.

when both words are out of dictionary, and if the edit distance[6] for u and v is smaller than 4, the edition is categorised as "probable misspelling" and the process is stopped.

Grammatical errors The use of a lemmatiser enables the analysis of real-word editions and their classification into one of the more specific grammatical error types: inflection, syntactic or semantic.

The examples of inflectional errors are of prime relevance as they are very frequent in languages with rich morphology, e.g. a grammatical gender disagreement:

- *System of a Down jest pierwszą grupą, która dwa razy w ciągu jednego roku (miał→miała) dwa albumy na szczycie.* [*System of a Down is the first group which ({he→she} has) two albums at the top of the chart in one year.*]

Even if their correction is an area of interest of current research [4], there is no tool that would handle the problem effectively. This kind of error indicates that u and v have equal lemmas. But it cannot be ascertained for sure whether the correction is a grammar or only style-related. The greatest confusion is about verbs differing only in tense or aspect, or nouns with the only change in the number, so we labelled all of them separately, e.g.:

- *Każdy odcinek (trwa→trwał) około pół godziny.* [*Each episode (takes→took) about half an hour.*]
- *Energie mają wyznaczone (miejsce→miejsca) w widmie elektromagnetycznym.* [*Energies have designated (place→places) in electromagnetic spectrum.*]

Editions in which u and v have different lemma are likely to be (1) a syntactic error correction if u and v belong to different grammatical classes and (2) semantic ones in the other case. Like in the case of inflectional errors, also the semantic errors which differ only in degree or aspect are classified separately. What is more, some editions recognised as semantic can be structural or pragmatic error corrections (according to Kukich classification [5]) and their automatic detection is probably impossible, e.g.:

- *Armia straciła ok. 1000 czołgów i (samolotów→samochodów) pancernych.* [*The army had lost about 1 000 tanks and armoured (planes→cars).*]

Other types of errors that are detected at this stage are insertions, deletions or substitutions of prepositions, pronouns and conjunctions.

Style errors Editions within abbreviations and acronyms are captured with the additional information provided by a lemmatiser. Other style adjustment editions are recognised by a thesaurus[7]. Using it before the grammatical errors detection prevents classifying style errors in the short edit distance into wrong category.

[6] As an edit distance we chose Damerau-Levenshtein distance.

[7] `http://synonimy.ux.pl/`

2.2 Conditions of acceptance

We allow at most one discarded edition in a sentence with the exception that the remaining editions are recognised as any type of the grammatical error. This is based on the observation that some editions (u_i, v_i) may be dictated by other editions (u_j, v_j) in the same sentence as in the case of inflection changes dictated by rewriting parts of a sentence, e.g.

- *Arytmetyka (jest→—) (najstarszą→najstarsza) i najbardziej (podstawową→ podstawowa) (gałęzią→gałąź) matematyki.* [*Arithmetic (is→—) the oldest and most elementary branch of mathematics.*]

The above sentence is rejected because the edition (*jest*, —) is not recognised with our heuristics and the rest of editions are grammatical error corrections. There is no such assumption in the case of spelling errors.

In order to avoid the situation when an edition is reverted (once or more times), for a given sentence with editions (u_i, v_i) we also check backward if the previously collected sentences from the same article are equal up to reversed editions (v_i, u_i). These sentences get cancelled altogether. It may imply an act of vandalism, or only a hesitation or differences in opinion of the editors, nevertheless, we do not take into account the editions of controversial words.

Finally, we perform post-processing included the entire sentence, as during the error recognition process we did the local analysis without consideration for a wider context. This stage has an impact on editions mainly concerning punctuation and latter case modification. For instance, we reject sentences for which the only change is: (1) deletion of a full stop from the end, (2) addition of a colon at the end, or (3) the conversion the first letter to lower case. We also discard text fragments addressing Wikipedia-specific content.

2.3 Difficulties

One of the main difficulties we had to deal with were the changes made by vandals [2], which was due to the edition of Wikipedia content being freely available to everybody. Vandalism usually involves minor changes that can be classified according to our heuristics as language error corrections.

The problem is solved on three levels:

1. Some acts of vandalism are reverted in the revision process and a relevant comment is added. So the revision with such comment (like *cancelled edits*, *revert after vandalism*, etc.) and the first previous editing of non-logged user are omitted.
2. Editions that include a vulgarisms or popular internet acronyms are rejected.
3. The backward checking is done as described in the previous section.

Another issue is an automatic edition done by Wikipedia's robots[8] mainly concerning data format, common abbreviations and the replacement of some HTML entities. From our point of view it is not, however, that relevant who made the correction, but that the error had occurred.

[8] Editions done by robot are easily detectable through *username* XML attribute.

3 Error corpus

We have applied proposed method to the Polish Wikipedia revision history creating the PlEWi corpus of naturally-occurring language errors and their corrections[9]. The dump of Wikipedia[10] contains 1,747,083 pages with about 910,000 articles.

The number of extracted text fragments with at least one potential language error is 1,532,275, including 1,303,806 (85.1%) of well-formed sentences. By “well-formed sentence” we mean each text fragment beginning with a capital letter, a number or quotation or hyphen mark and ending with the regular sentence delimiter: *.;?!”*. The remaining text fragments (228,469) are phrases, texts from tables, picture descriptions etc. About 23.0% of all the editions comes from anonymous users which may confirm a rather low risk of vandalism.

The total number of collected editions is 1,713,835, including 157,043 (10.9%) editions labelled as “probable misspelling”, i.e. words not existing in dictionary, but for which edit distance is smaller than 4. 70,828 corrections (0.05%) concern multi-word editions, whereas the number of deletions and insertions is 33,279 and 34,339, respectively (both constitute 0.02%). Detailed distribution of extracted error types is listed in Table 1.

Table 1. Error frequencies in PlEWi corpus.

Category	Error type	#
simple	punctuation	308,802
	case modification	220,533
	separable and inseparable writing	13,782
spelling	modification of diacritics (contextual)	241,777 (39,529)
	spelling	356,762
grammar	inflection (tense or number)	164,659 (43,340)
	syntactic	19,443
	semantic (aspect or degree)	64,600 (13,757)
	pron., prep., conj., particle-adverb	94,578
style	synonym	29,596
	abbreviation (year or age)	38,812 (21,431)
	“probable misspelling”	157,043

The number of corrections concerning diacritic modification together with the number of detected grammatical errors (but without the most doubted ones involving only of aspect, tense or number modification) can be considered as the total number of real-word errors. If editions identified as spelling error corrections would be considered as

[9] Corpus in YAML format and scripts, including a detailed documentation of presented heuristics, are publicly available at `http://www.staff.amu.edu.pl/~romang/wiki_errors.php`.

[10] The XML file of about 330 GB size from 14th July, 2012.

the non-word errors, then about 29.3% of all of them would be real-word. Including editions labelled as “probable misspellings” to non-word errors the ratio decreased to 24.1%. To the best of our knowledge it is the first estimation of this factor for the Polish language, when for English the range between 25% and 40% is used for the current research [5]. The lower relative number of real-word errors may be due to the fact that the average length of words in Polish is larger than in English.

4 Evaluation

To evaluate the effectiveness of the presented method for the language errors extraction and the quality of PlEWi corpus itself, we have manually checked 200 random text fragments from each category. For each first edition in each example we verified whether it is the right mark of the error type or not — it means that we calculated the precision value. The results are presented in Table 2.

Table 2. The evaluation of the selected error categories in PlEWi corpus

Category	#	Overall precision	#	Well-formed sentences
simple	200	0.86	146	0.90
spelling	200	0.98	173	0.99
grammar	200	0.73	170	0.71
style	200	0.99	169	0.99
probable misspelling	200	0.86	164	0.87

In the “simple” category (i.e. editions which recognition as error correction did not require any NLP tool) there were 14% of not valuable error corrections. Most of them were a faulty letter case modification and wrong insertion of comma, and next entire text fragments were syntactically incorrect as they probably came from a paragraph header or a bulleted list.

The application of the spell checker can explain a high precision value (98%) for recognition of spelling error corrections, and the lack of intentionally wrong editions.

On the other hand, 19% of editions among grammatical error corrections were connected only with the style improvement, such as grammatical aspect modification or the updating of the tense of some verbs (for instance, because time reference has been changed). The next 7% of them were pragmatic error corrections or context was not enough to decide if the edition was necessary. But most of these editions are marked by our heuristics separately, and after removing them from evaluation data set, the precision value increases to 0.80. In general, only 4 (2%) of all editions in this category result in an error form of word.

23% of editions which have not been recognised directly as any error correction were neither spelling nor grammatical error correction. But only in 10 (5%) cases an incorrect word was replaced by another incorrect form. As many as 112 (56%) editions concern the correction of named entities, 35 (18%) of them are inflectional error corrections, 11 (6%) stylistic changes, and 7 (4%) corrections of English or German word. Finally, 31 (16%) of them were proper spelling error corrections but concern less common or technical words.

The average precision value is 0.88. We do not present the recall value because the calculation of it would require a manually annotation of a quite large part of Wikipedia history.

5 Summary

In this paper, we presented automatic methods for the collection of a wide range of language errors and their corrections from histories of document editions. By applying them to the Polish Wikipedia revision history, we created the PlEWi corpus containing about 1.7 million naturally-occurring errors, including above 160 thousands of inflectional errors. As evaluation shows, the corpus is characterised by a high reliability of spelling error annotations (98.0%) and quite high for grammatical errors (72.5%).

We hope that the PlEWi corpus will become an important resource for developing and evaluating proofreading techniques for Polish.

References

1. Bušta, J., Hlaváčková, D., Jakubíček, M., Pala, K.: Classification of errors in text. In: RASLAN 2009 : Recent Advances in Slavonic Natural Language Processing. pp. 109–119 (2009)
2. Chin, S.C., Street, W.N., Srinivasan, P., Eichmann, D.: Detecting wikipedia vandalism with active learning and statistical language models. In: Proceedings of the 4th Workshop on Information Credibility. pp. 3–10 (2010)
3. Graliński, F., Jassem, K., Junczys-Dowmunt, M.: PSI-Toolkit: Natural language processing pipeline. Computational Linguistics - Applications pp. 27–39 (2012)
4. Kapłon, T., Mazurkiewicz, J.: The method of inflection errors correction in texts composed in polish language — a concept. In: Proceedings of the 15th international conference on Artificial neural networks — Volume Part II. pp. 853–858. ICANN'05, Springer-Verlag (2005)
5. Kukich, K.: Techniques for automatically correcting words in text. ACM Comput. Surv. pp. 377–439 (1992)
6. Leacock, C., Chodorow, M., Gamon, M., Tetreault, J.: Automated Grammatical Error Detection for Language Learners. Morgan and Claypool Publishers (2010)
7. Max, A., Wisniewski, G.: Mining naturally-occurring corrections and paraphrases from wikipedia's revision history. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (2010)
8. Miłkowski, M.: Automated building of error corpora of polish. In: Corpus Linguistics, Computer Tools, and Applications State of the Art. PALC 2007, pp. 631–639. Peter Lang (2008)
9. Zeng, H., Alhossaini, M.A., Ding, L., Fikes, R., McGuinness, D.L.: Computing trust from revision history. In: Proceedings of the 2006 International Conference on Privacy, Security and Trust (2006)
10. Zesch, T.: Measuring contextual fitness using error contexts extracted from the wikipedia revision history. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 529–538 (2012)