Bootstrapped OCR error detection for a less-resourced language variant

Adrien Barbaresi
Berlin-Brandenburg Academy of Sciences & Austrian Academy of Sciences
barbaresi@bbaw.de

Abstract

This study focuses on isolated error detection in a retro-digitized newspaper corpus published from 1946 to 1990 in the former German Democratic Republic. As there are OCR errors throughout the corpus but no clean reference for this variant of German, automatic OCR correction implies to overcome data sparseness and non-standard spelling, including compounds and inflected forms. The contributions of this paper are (1) a method to bootstrap detection of potential misspellings, (2) an assessment of several types of training data, and (3) an evaluation of several off-the-shelf candidate selection techniques. The chosen solution based on statistical affix analysis reaches an accuracy 10 points higher than existing morphological analysis systems on error detection, while a combination of fuzzy and approximate string search performs best for error correction. The criteria are met since it is possible to correct erroneous tokens without introducing too much noise.

1 Introduction

The study presented in this paper stems from a collaboration with historians to work on a diachronic newspaper corpus published in and at the time of the former German Democratic Republic (GDR/East Germany). The corpus has been digitized by a library consortium with limited resources, and the advertised quality is 95% error-free content. While no precise unit is given, it can be assumed this is on character level, which could qualify as average optical character recognition (OCR) accuracy (Holley, 2009), and which also leaves much room for improvement on token level. Numerous OCR errors can be expected throughout the texts, i.e. neither author ignorance, nor typographical errors on typing, but transmission and storage errors (Peterson, 1980).

To reduce the error rate, automatic post-processing of digitized documents is necessary. As the retro-digitized newspaper (*Neues Deutschland*) is a first attempt to grasp language use in the GDR on a large scale, there are no available corpora of this kind to train statistical models on or evaluate the results, although commonly used noisy channel models (Brill and Moore, 2000) work best on manually corrected training data, and system evaluations are performed on series of string pairs (Eger et al., 2016). In the absence of a gold standard, a bootstrap method has to be found in order to predict errors accurately without a reference. The overall precision has to be high, otherwise the correction process could degrade the corpus more than it improves it.

I focus on non-word misspellings, strings that are not found even in a large dictionary (Flor, 2012), and I develop a corrector, which implies detecting misspelled words and trying to find the most likely correct word (Peterson, 1980). This has to be done on a single OCR output, methods based on different OCR engines (Klein and Kopel, 2002) are not applicable. The contributions of this paper are as follows: (1) a corpus-based method to bootstrap detection of potential misspellings; (2) an assessment of several types of training data; and (3) an evaluation of several off-the-shelf candidate selection techniques.

2 Problem description

2.1 Error detection task

In the remainder of this article, emphasis lies on isolated non-word error correction (Kukich, 1992), also known as type-wise canonicalization techniques (Jurish, 2010) and single-token non-word OCR error correction using non-contextual algo-

rithms (Flor, 2012). Word segmentation issues are existent, but the way there are processed by such a component as well as others in a classical annotation toolchain is too difficult to benchmark, so that they have to be addressed separately. The problem tackled in this article can be split into three tasks: detection of an error, generation of candidate-corrections, and ranking of the corrections (Kukich, 1992).

Since progresses in hardware have been significant since the 90s, it is now possible to design an "ideal system" which involves "broad lexical coverage" and a lexicon as large as 100,000 words (Kukich, 1992). The task can be performed using a large database of token n-gram occurrences (Carlson and Fette, 2007). However, the context of this study is far from the "idealized conditions" described by Généreux et al. (2014), i.e. no more than two edit operations and a perfect dictionary. There are indeed substantial problems with error models driven by rules when the Levenshtein distance (Levenshtein, 1966) between error and correct string is higher than 2, and the least distant string is not necessarily the best candidate.

Since there is no proper dictionary to derive all correct word forms from, the task cannot be reduced to a normalization of out-of-vocabulary tokens to an in-vocabulary standard form, as commonly formulated (Han et al., 2013). More specifically, due to the diversity of morphology and flexion in German, rare forms potentially unknown to dictionaries may be correct (e.g. *Leninschem*, dative form "relative to Lenin", or *Spitzenlastfahrweise*, a technical term used for power plants), and keeping case markers intact is paramount.

Following from the differences listed above, the task differs from classical OCR-post-correction processes in the way that the tokens to be corrected are partly divergent but fully correct utterances, and partly OCR-related errors. The ratio between them is expected to be 95 to 5%, but it is impossible to assess with precision and it varies in time. In that sense, it is comparable to normalization of short text messages in that lexical variants may be intentionally generated (Han et al., 2013), and my goal is to overcome data sparseness.

2.2 Related results

Benchmarks are hard to come by since to my best knowledge there is no quantitative study on spell-checking for texts published in the GDR. Several methods tested on English in a seminal article (Kukich, 1992), with comparatively small dictionaries, yield top accuracies between 0.75 and 0.81. Regarding inflected languages, the TISC system for Dutch advertises a precision of 0.60, a recall of 0.67, and an F-measure of 0.63 on diachronic newspaper corpora (Reynaert, 2004), while its successor TICCL achieves a precision of 0.926, a recall of 0.894, and a F-measure of 0.910 when used without lexicon on contemporary parliament acts (Reynaert, 2011). Concerning language variants, character-level models on Egyptian Arabic dialect reach an accuracy of 0.805 on out-of-vocabulary and 0.946 on in-vocabulary words (Farra et al., 2014).

2.3 Characteristics of the corpora

The *Neues Deutschland*-corpus (ND) spans practically the time of existence of the GDR: it comprises 1.46 million articles published from 1946 to 1990, and about 444 million tokens in total. Its OCR quality varies significantly due to font changes and apparently uneven digitization.

To build a reference, two comparable corpora in size and time span are taken into consideration. Both were published in the Federal Republic of Germany (West Germany): (1) *Die Zeit* (*DZ*; 1946-2015; 1.12 million articles; 529 million tokens), and (2) *Der Spiegel* print edition (*SP*; 1947-2015; 324,000 articles; 246 million tokens). These corpora have been crawled from online archives, digitization has been undertaken by the publishers; the documents used to build a corpus are thus natively digital and they are practically exempt of OCR-related errors.

Comparison on type level shows significant discrepancies between the newspaper corpora, with a higher absolute number of types for ND, and low overlapping between the types: only 23.4% of ND's alphabetic types are found in a combination of DZ and SP. This indicates that while errors may have been contained on character level, the dispersion on type level is very high, meaning that there are a relatively high number of erroneous variants for each potential error-free token, and that dictionary coverage is low in any case.

2.4 Linguistic setting

Additionally, there are peculiarities of German as spoken in the GDR which need to be clarified. The newspaper uses a written standard so that in general no dialectal/regional variance is to be expected. However, there are a number of differences regard-

ing institutions, social roles, and words used in everyday life. This is particularly true for compound names, due to the flexibility of German: between both sides of the boundary, a high number of true lexical differences are to be found in (1) comparatively unusual but frequent compounds (e.g. *antiimperialistisch*, anti-imperialistic), (2) roots and compounds typical for systemic differences (e.g. *Kombinat* for business group or conglomerate in East Germany), and (3) rare compounds due to the focus on particular aspects (e.g. *Euterkontrolle*, udder control).

Proper nouns are also potentially an issue because of the diverging national and ideological references. Nonetheless, the difference seems to be of quantitative nature, since most person and place names used in the East appear in the West, albeit with a much lower frequency. This discrepancy indicates that frequency information in reference corpora may not be significant.

3 Method

The overlap between reference and correction corpora is low, so that working on improving dictionary coverage may not be the best approach. I use a corpus-based morphological analysis to find potential OCR-errors, whereas approximate matching (Hall and Dowling, 1980) and fuzzy search algorithms (Hauser et al., 2007; Généreux et al., 2014) based on character n-gram models are used to generate candidates for replacement and find the best one.

3.1 Error detection

Morphological analysis in German is performed by software such as SMOR (Schmid et al., 2004), which is suitable for texts of this period due to its training materials. It is expected that since it somehow reflects the logic of the language, it does not output any analysis for words which do not exist, whereas it would do so for rare compounds and even proper nouns.

The method introduced here is data-driven and grounds on affix analysis (Peterson, 1980). Relevant information is stored in a trie (Fredkin, 1960), a data structure allowing for prefix search and its reverse opposite in order to look for sublexicons, an approach used for instance in the case of agglutinative languages (Agirre et al., 1992). Compound splitting is highly necessary in morphologically rich languages (Reynaert, 2004), tokens are decomposed whether thex contain hyphens or not. The smallest possible token length for learning and searching is fixed to 4 characters. The affix and morpheme trees are learned from a types list. Simple rules are added to account for joins between compounds as well as inflection-related endings (-s, -en, etc.) in order to cope with rare phenomena which might not be present in the training data. The detection algorithm consists of one or two iterations of a search for the longest prefix and suffix as well as sanity checks to see if the rest could itself be an affix or a word of the dictionary.

3.2 Candidate selection

Candidates are found and ranked using bigram and trigram similarity (Zamora et al., 1981). On top of the similarily, fuzzy string matching already used for spelling-correction in historical texts (Hauser et al., 2007; Généreux et al., 2014) as well as approximate string matching are used. The approach tends to be conservative, nothing is modified if nothing is found within the bounds of a search space. Moreover, the agreement between both search algorithms is also evaluated. To account for inflexions, endings are normalized to the form of the original token in case a correction is suggested; capitalization is also restored to the original state.

4 Results

4.1 Evaluation data

The data for this experiment consist of a "difficult but realistic" (Kukich, 1992), "clear" set of string pairs, some misspellings and some correct but rare types; it contains a fair proportion of proper nouns as well as shorter items. The candidates have been found using frequency lists and morphological analysis tools, the list is designed to be difficult for the tools at hand. For the sake of evaluation, all cases can be considered to be unambiguous.

There are 500 non-word errors with corrections, with a Levenshtein distance comprised between 1 and 5 (mean 1.7, standard deviation 0.8): *Kriegsvqrhereitung*, *Sdiwermasdiinenbau*, *Tsdiiangkaischek*. On the other hand, there are 500 rare but correctly spelled words including inflected forms for the detection of false positives: *Komsomolzen*, *Plastfolie*, *Antiimperialistischen*, *CSSR-Mädchen*, *Kleinstübertrager*, etc. The dataset is available online.[1]

[1] http://clarin.bbaw.de/de/objects/dwds:7/

	Voc. size	Precision	Recall	F-score	Accuracy
Spellchecker					
hunspell (*de_DE*)	∼ 75,000	.583	1	.737	.643
Morphological analysis (no result for the word)					
ZMORGE	∼ 78,000	.630	.926	.750	.691
MORPHISTO	∼ 18,200	.638	.948	.763	.705
SMOR	∼ 50,000	.701	.946	.805	.771
Affix tries and composition rules					
Top-10% ND	725,995	.806	.406	.540	.654
Top-10% DZ+SP	596,984	.797	.924	.856	.844
Top-10% WEB	2,205,332	.855	.846	.850	.851
Intersection DZ+SP+KERN	757,953	.842	.904	.872	.867
Top-35% KERN	814,156	.837	.914	.874	.868
Top-10% DZ+SP+KERN	897,359	.842	.908	.874	.869
Intersection DZ+SP	1,620,976	.866	.890	.878	.876

Table 1: Evaluation of several error detection strategies, ordered by ascending accuracy

4.2 Error detection

I resort to morphological analysis to see if the words are to be corrected or not, the results are summarized in Table 1. My evaluation features the Enchant interface to the hunspell spell-checker[2] (de_DE-locale), common morphological analysis software such as *Morphisto* (Zielinski et al., 2009), *SMOR* (Schmid et al., 2004) and its enriched version based on the Wiktionary *Zmorge* (Sennrich and Kunz, 2014). The models used are the standard off-the-shelve ones, since no training material is available for the texts, and since standard training is assumed to be close enough to newspaper text.

My method uses affix trees induced as described above from West-German newspaper texts, on tokens with a minimum length of 4 characters. Additionally, the DWDS core corpus (Geyken, 2007), a balanced corpus for German in the 20th century (*KERN*; 1900-1999; 123 million tokens) is taken as an error-free reference. As it has been shown that web corpora could lead to better OCR correction (Strohmaier et al., 2003; Whitelaw et al., 2009), results based on frequent word forms extracted from a giga-token "clean" web corpus of German (Barbaresi, 2016) are referenced in the benchmark (*WEB*; 2002-2015; 2.1 billion tokens), although the corpus is neither geographically nor topically focused.

The results show that the efficiency of detection does not rely primarily on vocabulary size, the training corpora are preponderant for all tested solutions. The method introduced in this article works best in terms of precision, F-score, and accuracy, albeit with vocabularies sizes ten to twenty times larger than other tools. It cannot be trained on noisy corpus data since even a frequency filter cannot eliminate all OCR errors in the training with ND. Manual screening confirms that there are errors to be found in the top-10% of ND types, showing the extent of the problem to be treated.

Clean contemporary data from the DWDS-core corpus achieve good results even if the frequency range taken for the study is stretched toward less frequent types. The types extracted from the Web corpus are not optimal: since it does not cover the right text type and the right period, much more information is needed to achieve a similar result, thus introducing more noise. However, corpus size is not an issue with web corpora, and the results still are a positive indication as to their usefulness for general purposes, with a well-balanced ratio between precision and recall. The affix models based on contemporary West-German newspapers (DZ and SP) generally achieve better results; training data featuring not a frequency filter but an intersection (types present at least once in both newspapers) seem to eliminate potential noise due to *hapax legomena* while gathering enough information to provide a small boost concerning accuracy.

The output of morphological analysis based on the top-10% types of DZ+SP is used to discriminate between the tokens in the benchmark, since my method and this dataset provide the best F-measure as well as the best accuracy.

[2]http://www.abisource.com/projects/enchant/

Algorithm	Prec.	Rec.	F-sc.	Acc.
Approximate	.942	.524	.674	.746
Fuzzy	.922	.594	.723	.772
Combination	.949	.524	.675	.748

Table 2: Evaluation of error correction algorithms

4.3 Candidate selection

Due to the configuration of the data set the search space is limited to a maximum Levenshtein distance of 5. Whether candidates are ranked by distance or by frequency does not make noticeable changes because the algorithms already use a frequency measure internally. The parametrization of character n-grams does not bring a significant boost either: 2- or 3-grams achieve similar results. Punctuation and flexion rules yield small improvements. To replicate the results in order to make sure that no artifacts arise from a particular algorithm implementation, the method has been tested in Perl and Python using corresponding modules and packages[3], with similar results.

Due to the data used the maximum recall is 0.89. The results are summed up in Table 2. Approximate string search yields the best results in terms of precision while the fuzzy string search algorithm performs better in terms of recall, F-score, and accuracy. The best conservative approach seems to be a combination of fuzzy set and approximate string search (intersection). Although the recall values are low (between 50 and 60%), the accuracy on out-of-vocabulary tokens slightly falls short to the results of Farra et al. (2014) for Egyptian dialect, and this first experiment already meets the criteria for text correction, since erroneous tokens would be corrected without introducing too much noise.

Regarding qualitative evaluation, frequency-based error correction such as *Usa-Ausbeuter* (US-exploiter) in *Usa-Aushelfer* (US-aides, rare and generally used in a military context) would be grammatically correct but completely wrong as far as historical analysis is concerned. However, most recurring errors are of secondary importance as they deal with specialization (*Radialbohrmaschine* erroneously changed to *Spezialbohrmaschine*), or evolving normalization of proper nouns across time (*Bjelorußland* and *Belorußland*).

A way to address the mistakes may be to perform a proper candidate re-ranking (Flor, 2012), for instance by changing the costs for Levenshtein distance calculation (Hauser et al., 2007). First tests show two difficulties due to discrepancies and inflected forms: either the solution is not even in the candidate list or the distance costs do not perform evenly.

[3]Python: *marisa-trie*, *fuzzyset*, and *ngram* modules.
Perl: *Tree::Trie*, *Text::Fuzzy*, and *String::Approx*.

5 Conclusion

I have provided a method to bootstrap detection of potential misspellings in a language variant without existing standard data. Concerning error detection, morphological analysis trumps out-of-vocabulary methods as well as regular spell-checkers. Additionally, statistical affix analysis trumps morphological analysis, with accuracies up to 10 points higher than SMOR. Clean and if possible contemporaneous corpus data make a positive difference in the benchmark, and although GDR-specific vocabulary is rare in web corpora they seem to have potential as a supplementary resource. Error correction is best performed by a combination of off-the-shelf candidate selection techniques, in order to find the right balance between statistical and rule-based approaches. In both cases, results are in line with the criteria for the task, since they would correct erroneous tokens without introducing too much noise.

Acknowledgments

This work has been supported by a CLARIN-D special interest group dedicated to late modern and contemporary digital history (*FAG-9*).

References

Eneko Agirre, Inaki Alegria, Xabier Arregi, Xabier Artola, A Díaz de Ilarraza, Montse Maritxalar, Kepa Sarasola, and Miriam Urkia. 1992. XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology. In *Proceedings of the 3rd conference on Applied Natural Language Processing*, pages 119–125. Association for Computational Linguistics.

Adrien Barbaresi. 2016. Efficient construction of metadata-enhanced web corpora. In *Proceedings of the 10th Web as Corpus Workshop*, pages 7–16. Association for Computational Linguistics.

Eric Brill and Robert C Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 286–293.

Andrew Carlson and Ian Fette. 2007. Memory-based context-sensitive spelling correction at web scale. In *Proceedings of ICMLA*, pages 166–171. IEEE.

Steffen Eger, Tim vor der Brück, Alexander Mehler, et al. 2016. A Comparison of Four Character-Level String-to-String Translation Models for (OCR) Spelling Error Correction. *The Prague Bulletin of Mathematical Linguistics*, 105(1):77–99.

Noura Farra, Nadi Tomeh, Alla Rozovskaya, and Nizar Habash. 2014. Generalized Character-Level Spelling Error Correction. In *Proceedings of the Annual Meeting of the ACL*, pages 161–167.

Michael Flor. 2012. Four types of context for automatic spelling correction. *TAL*, 53(3):61–99.

Edward Fredkin. 1960. Trie Memory. *Communications of the ACM*, 3(9):490–499.

Michel Généreux, Egon W Stemle, Verena Lyding, and Lionel Nicolas. 2014. Correcting OCR errors for German in Fraktur font. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*, pages 186–190. Pisa University Press.

Alexander Geyken. 2007. The DWDS corpus: A reference corpus for the German language of the 20th century. *Collocations and idioms: Linguistic, lexicographic, and computational aspects*, pages 23–40.

Patrick AV Hall and Geoff R Dowling. 1980. Approximate String Matching. *ACM computing surveys (CSUR)*, 12(4):381–402.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):5.

Andreas Hauser, Markus Heller, Elisabeth Leiss, Klaus U Schulz, and Christiane Wanzeck. 2007. Information access to historical documents from the Early New High German period. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Rose Holley. 2009. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3/4).

Bryan Jurish. 2010. More than words: using token context to improve canonicalization of historical German. *JLCL*, 25(1):23–39.

Shmuel T Klein and Miri Kopel. 2002. A voting system for automatic OCR correction. In *Proceedings of the Workshop on Information Retrieval and OCR at SIGIR*, pages 1–21.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4):377–439.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 8, pages 707–710.

James L Peterson. 1980. Computer programs for detecting and correcting spelling errors. *Communications of the ACM*, 23(12):676–687.

Martin Reynaert. 2004. Multilingual text induced spelling correction. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pages 117–124. Association for Computational Linguistics.

Martin Reynaert. 2011. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition (IJDAR)*, 14(2):173–187.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. In *LREC*.

Rico Sennrich and Beat Kunz. 2014. Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of LREC*, pages 1063–1067. ELRA.

Christian M. Strohmaier, Christoph Ringlstetter, Klaus U Schulz, and Stoyan Mihov. 2003. Lexical postcorrection of ocr-results: The web as a dynamic secondary dictionary? In *Proceedings of ICDAR*, pages 1133–1137.

Casey Whitelaw, Ben Hutchinson, Grace Y Chung, and Gerard Ellis. 2009. Using the web for language independent spellchecking and autocorrection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 2, pages 890–899. Association for Computational Linguistics.

EM Zamora, Joseph J Pollock, and Antonio Zamora. 1981. The use of trigram analysis for spelling error detection. *Information Processing & Management*, 17(6):305–316.

Andrea Zielinski, Christian Simon, and Tilman Wittl. 2009. Morphisto: Service-oriented Open Source Morphology for German. In *State of the Art in Computational Morphology*, pages 64–75. Springer.