

Lexical Postcorrection of OCR-Results: The Web as a Dynamic Secondary Dictionary?

Christian Strohmaier
SfS – Univ. of Tübingen
strohmai@sfs.uni-tuebingen.de

Christoph Ringlstetter, Klaus U. Schulz*
CIS – Univ. of Munich
{kristof,schulz}@cis.uni-muenchen.de

Stoyan Mihov*
LPDP – Bulgarian Academy of Sciences
stoyan@lml.bas.bg

Abstract

Postcorrection of OCR-results for text documents is usually based on electronic dictionaries. When scanning texts from a specific thematic area, conventional dictionaries often miss a considerable number of tokens. Furthermore, if word frequencies are stored with the entries, these frequencies will not properly reflect the frequencies found in the given thematic area. Correction adequacy suffers from these two shortcomings. We report on a series of experiments where we compare (1) the use of fixed, static large-scale dictionaries (including proper names and abbreviations) with (2) the use of dynamic dictionaries retrieved via an automated analysis of the vocabulary of web pages from a given domain, and (3) the use of mixed dictionaries. Our experiments, which address english and german document collections from a variety of fields, show that dynamic dictionaries of the above mentioned form can improve the coverage for the given thematic area in a significant way and help to improve the quality of lexical postcorrection methods.

1. Introduction

Postcorrection of OCRed text is generally based on electronic dictionaries [5, 3, 8, 4, 1]. The relevance of the choice of the dictionary for correction accuracy is often stressed [7]. However, since scanned corpora often belong to specific thematic areas, general purpose dictionaries usually fail to reflect vocabulary and word frequencies of the text. Consider a corpus C^{ocr} obtained from an OCR-analysis of a printed version of the corpus C . The perfect dictionary D for postcorrection satisfies three principles: (1) D con-

tains each word of C , (2) D contains only words from C , and (3) for each word W , D stores the frequency of W in C . Principle (1) guarantees that in principle each garbled word of C^{ocr} can be properly corrected. (2) helps to avoid improper “corrections”, and (3) helps in the absence of better information to desambiguate between several correction candidates for a given garbled token.

In practice the perfect dictionary is not available. Emphasizing principle (1) it is sometimes recommended to use a large-scale dictionary which contains a maximal amount of common words, terminological expressions as well as proper names and abbreviations. Other sources recommend to use a dictionary that contains only the most frequent tokens, in order to find a compromise between principles (1) and (2). Word frequencies in dictionaries are usually obtained from an analysis of a large corpus, such as the Brown Corpus [2] or the British National Corpus (BNC).

The problem with these approaches is that the dictionary is not adapted to the given thematic topic. In practice, depending on the topic, C is likely to contain a nontrivial amount of tokens that are not found in D , even if D is very large. In addition, word frequencies of D will not match the word frequencies found in C . Correction adequacy suffers from these two shortcomings. In this paper we look at a simple

Hypothesis: *Most tokens that occur in a text of a given area can be found in web pages with a direct relationship to the given thematic field. Relevant web pages can be retrieved using simple queries to internet search machines. Analyzing the vocabulary of such pages, thematic dictionaries can be built automatically that improve the coverage of standard dictionaries in a significant way, yield estimates for the occurrence frequencies of words in the given area that are more reliable than frequencies derived from general purpose corpora and thus help to improve the correction adequacy of systems for lexical postcorrection. The*

*Funded by VolkswagenStiftung

negative effect of incorrect tokens that may occur in web pages is neglectable.

Note that evidence for this hypothesis would motivate research towards correction systems where appropriate software for the analysis of the vocabulary of web pages is fully integrated that can be used to dynamically derive in an “on-line” fashion domain specific dictionaries that are joined with static background dictionaries in a given application.

In order to test the hypothesis we considered a variety of specific thematic topics from distinct fields. After selecting english and german text corpora for each topic we made a series of experiments where large-scale conventional dictionaries for the given language, special dictionaries for proper names, geographic names, acronyms and abbreviations, dictionaries with most frequent words, as well as “dynamic” thematic dictionaries with web vocabulary of the kind described above were composed in different ways. For the sake of comparison, also the “perfect dictionary” of the underlying text (s.a.) was used. In order to judge the quality of each dictionary for lexical postcorrection, each test corpus was analyzed with commercial OCR-software. Output files were corrected, following a simple model for lexical postcorrection and using the given dictionary. We then calculated several parameters that are relevant for automated correction system (correction accuracy, s.b.) and for interactive correction systems (e.g., false friend rate, no chance rate, inspection rate, s.b.).

Our results, which are described in Section 4, show that dynamic dictionaries of the above mentioned form can improve the coverage for the given thematic area in a significant way and help to improve the quality of lexical postcorrection methods.

2. Evaluation parameters and correction model

In our experiments, lexical coverage of a dictionary D is measured using the original corpus C . Hence it is independent from OCR-recognition results and from correction strategies. We distinguish between “normal” tokens of C , which are composed of standard letters only, and “abnormal” tokens including other symbols (e.g., “thaw_request”, “#0-358-81160-1”). Tokens of the latter type are usually not collected in dictionaries. Hence we define **lexical coverage** of a dictionary D as the percentage of normal tokens of C that occur in D .

In order to judge the quality of a given dictionary D for postcorrection of OCR-results the given corpus C was printed, copied, scanned and analyzed with commercial OCR-software. In this way we obtained a parallel corpus C^{ocr} . We then used a simplified model¹ for lexical post-

correction of C^{ocr} . In the sequel, W^{ocr} denotes a token of C^{ocr} and W^{pc} denotes the correction result.

1. Fix an upper bound b_0 for the length-sensitive Levenshtein distance² $d'(W_{cand}^{pc}, W^{ocr})$ between a token W^{ocr} and a correction candidate W_{cand}^{pc} and a threshold f_0 for the frequency of W_{cand}^{pc} .
2. For each abnormal token W^{ocr} define $W^{pc} := W^{ocr}$.
3. For each normal token W^{ocr} occurring in D define $W^{pc} := W^{ocr}$.
4. If a normal token W^{ocr} of C^{ocr} is not found in D , compute all entries V of D such that the Levenshtein distance $d(V, W^{ocr})$ (s. footnote) is minimal w.r.t. all entries of D . Among all entries with minimal distance, let W_{cand}^{pc} denote the one with the highest frequency f_{cand} . If $d'(W_{cand}^{pc}, W^{ocr}) \leq b_0$ and $f_{cand} \geq f_0$, then define $W^{pc} := W_{cand}^{pc}$, else define $W^{pc} := W^{ocr}$.

After an automated alignment of C and C^{ocr} , splits and merges of tokens were filtered out. We also excluded tokens of C (if any) with obvious spelling errors. Each remaining token W^{ocr} of C^{ocr} corresponds to a unique token W of C . We say that W^{ocr} is *properly corrected* iff $W^{pc} = W$.

In the present context, abnormal tokens of C^{ocr} are not subject to lexical correction. Hence we define **correction accuracy** as the percentage of normal tokens of C^{ocr} that are properly corrected using D . In order to simplify comparison with correction accuracy, **OCR-accuracy** is also measured with respect to *normal* tokens only and defined as percentage of normal tokens of C^{ocr} representing correct recognition results. Note that the above notion of correction accuracy depends on the bounds b_0 and f_0 that are introduced in Step 1 of the correction model. In our experiments we computed the bounds that lead to optimal correction accuracy. As a matter of fact, in a practical application optimal bounds can only be estimated on the basis of training data or partial evaluations. Our concern is a comparison of dictionaries, hence for simplicity we used optimal bounds.

Several reasons may exist that a token of C^{ocr} is not corrected properly. In order to have a clearer picture on the influence of the size of the dictionary we considered all triples of the form (W, W^{ocr}, W^{pc}) where $W \neq W^{pc}$. Since abnormal tokens are not subject to lexical correction we ignored triples where W^{ocr} is abnormal. The remaining postcorrection errors were classified in the following way:

if ed and refi ned in many different ways. Since we just want to judge the quality of dictionaries, we adopted a general and simple model.

²The *standard Levenshtein distance* [6] between words W and V , denoted $d(V, W)$, is the minimal number of letter insertions, deletions and substitutions that are needed to transform V into W . The *length-sensitive Levenshtein distance* is $d'(V, W) := d(V, W)/(|V| + |W|)$ where $|U|$ denotes the length of U .

¹Clearly, in a realistic application our correction model could be mod-

1. “false friends”: $W^{ocr} \in D$. Here $W^{pc} = W^{ocr}$.
2. The OCR-result is actively corrected, with wrong result ($W^{ocr} \notin D$ and $W^{pc} \neq W^{ocr}$). We distinguish three subcategories, (a) “wrong candidate” errors where $W \in D$ (here $W \neq W^{ocr}$), (b) “infelicitous correction” errors where $W \notin D$ and $W = W^{ocr}$, and (c) “no chance I” errors where $W \notin D$ and $W \neq W^{ocr}$.
3. The OCR-result is left unmodified, with wrong result ($W^{ocr} \notin D$ and $W^{pc} = W^{ocr}$). We distinguish three subcategories, (a) “too cautious” errors where $W_{cand}^{pc} = W$, (b) “wrong candidate and bound” errors where $W_{cand}^{pc} \neq W$ and $W \in D$, and (c) “no chance II” errors where $W_{cand}^{pc} \neq W$ and $W \notin D$.

“False friend” errors can only be avoided with a smaller dictionary. “No chance” errors can only be avoided with larger dictionaries. The **false friend rate** is defined as the number of false friends divided by the number of normal tokens of C^{ocr} . **No chance rate** is defined accordingly. For the other errors mentioned above, the size of the dictionary is less influential.

In a simplified scenario for *interactive OCR-correction* we might decide to inspect all pairs (W^{ocr}, W_{cand}^{pc}) where W^{ocr} is not in D .³ In order to measure the amount of work we define the **inspection rate** as the number of normal tokens of C^{ocr} that are not in D divided by the number of normal tokens of C^{ocr} .

3. Corpora and dictionaries

Specific thematic topics. The topics used for the tests can be found in Table 1.

Language alternations. In order to study the influence of the underlying language, all working steps and tests to be described below were carried out in two variants, respectively using English (E) or German (G) as the basic language. A serious problem for lexical analysis of german texts is the high amount of composed words. The number of composite nouns is not restricted, hence there is no way to build a complete dictionary for all compounds.

Parallel test corpora. For each subfield and language (E, G), an electronic test corpus C was collected with documents belonging to the respective area. Table 1 gives the number of tokens of each corpus. Each corpus C was printed, copied once, scanned and analyzed with industrial OCR-software to produce a parallel corpus C^{ocr} .

Static lexical resources. Five dictionaries, respectively containing conventional english/german words (D^E , D^G),

³Clearly, such a strategy can only be used if the user is willing to accept a certain number of false friends. In order to avoid false friends, the user has to inspect every normal token.

Topic	English Corpus	German Corpus
Botany	4.478	6.340
Neurology	6.801	5.691
Holocaust	7.545	5.595
Roman Empire	8.000	7.037
Fishes	10.375	7.719
Mushroom	7.561	6.143

Table 1. Specific topics and number of tokens of test corpora.

D^E	D^G	D_p	D_g	D_a
315.300	2.235.136	372.628	147.415	1.185

Table 2. Sizes of static subdictionaries.

international proper names (D_p), geographic names (D_g), and abbreviations (D_a) with frequency information were at our disposal. Frequencies were obtained via an analysis of a 2 TeraByte subcorpus of the WWW from 1999. A language classifier was used in order to evaluate english (D^E) and german (D^G) web pages only. The size (number of entries) of each component dictionary is given in Table 2. Note that already D^E and D^G are very large. Using these dictionaries we compiled four additional dictionaries. D^{EG+} (resp. D^{GE+}) is the union of all dictionaries mentioned above, frequencies based on english (resp. german) web pages. D_{\downarrow}^{EG+} and D_{\downarrow}^{GE+} respectively represent the 100.000 most frequent tokens of D^{EG+} in english and german web pages. The use of two languages in D^{EG+} (resp. D^{GE+}) is motivated by the german (english) expressions that are found in certain english (german) corpora.

Perfect dictionary. We computed for each corpus C the perfect dictionary as defined in the introduction. Correction results obtained with the perfect dictionary serve as an upper limit that cannot be improved.

Dynamic lexical resources. In order to create a *specific dictionary* for each subarea and language, a query with 25 terminological expressions automatically extracted from C was sent to the AllTheWeb internet search engine, together with the appropriate restriction on the language.⁴ 1000 top-scored web pages from the answer set were selected. The reachable pages were used to build a repository of texts. Analysing the vocabulary of the repository we derived a thematic dictionary with frequency information for each entry. In this way we obtained the domain specific dictionary.

⁴In order to avoid a situation where we accidentally reuse parts of the parallel corpus for retrieving the domain-specific dictionaries, a specific fingerprint was built for each page of the corpus. Each fingerprint was sent as a query to AllTheWeb, and all the answer documents were excluded from the dictionary construction process.

ies D^{EW} and D^{GW} . A primitive trick was used to obtain a better closure under inflectional variants. In the correction process we left a token W^{ocr} unmodified if W^{ocr} was identical to an entry of D^{EW} (resp. D^{GW}) modulo an inflectional suffix. As english (resp. german) inflectional endings we used the suffixes -, -ed, -s, -ing, -ly, -less, -er (resp. -, -e, -es, -r, -s, and -n following a vocal, r or l).

Mixed lexical resources. The dynamic dictionaries D^{EW} and D^{GW} were joined with D^{EG+} , obtaining D_{EG+}^{EW} and D_{EG+}^{GW} .

4. Results, Comments and Resumee

The results for english and german corpora are respectively collected in Tables 3 and 4. “No chance” errors and “false friend” errors are given in absolute numbers, all other entries represent percentages.

The *lexical coverage* of the small static dictionaries D_{EG+}^{EW} and D_{EG+}^{GW} is often weak. Remarkably, the coverage reached with the crawled dictionaries D^{EW} , D^{GW} alone is always better than the coverage of the maximal static dictionaries D^{EG+} , D^{GE+} . The coverage of the combined dictionaries D_{EG+}^{EW} , D_{GE+}^{GW} is always much better than the coverage reached with D_{EG+}^{EW} and D_{GE+}^{GW} . Hence, looking at coverage, D_{EG+}^{EW} and D_{GE+}^{GW} represent an optimal choice.

As to *correction accuracy*, optimal results are either obtained with combined dictionaries D_{EG+}^{EW} , D_{GE+}^{GW} or with crawled dictionaries. When we define the difference between OCR-accuracy and correction accuracy obtained with the perfect dictionary as the maximal improvement of accuracy that can be reached, the real improvement using the combined dictionary for english corpora is 25%, 22%, 16%, 78%, 60%, 58%, for german corpora we obtain 13%, 2%, 50%, 3%, 4%, 18%. For some of the german corpora, correction accuracy does not go much beyond plain OCR accuracy. It should be kept in mind that we use a very simple model for lexical postcorrection. Probably better results could be reached with more sophisticated models (e.g. [7]).

Not surprisingly, the number of *false friend errors* (resp. *no chance errors*) grows (decreases) with the size of the dictionary. Note that even for the perfect dictionary D , “no chance” errors may occur if an abnormal token W of C is recognized as a normal token W^{orc} since $W \notin D$ in this case. We found that a very large amount of false friend errors is caused by small tokens of length 1 – 3. For those words, dictionary lookup is not very selective because of many abbreviations etc.

The use of large dictionaries leads to a significant reduction of the *inspection rate*. Again optimal results are obtained with the combined dictionaries D_{EG+}^{EW} , D_{GE+}^{GW} . This shows that combined dictionaries are particularly interesting for interactive postcorrection.

Language differences are obvious. For corresponding topics and dictionaries, lexical coverage obtained for the german corpus is always lower than the coverage reached for the english corpus, due to composition of words in german language. Consequently, correction accuracy (inspection rate) obtained for english corpora is generally better than for german corpora. For english texts often small static dictionaries lead to better accuracy results than static dictionaries of a maximal size. In contrast, for german texts, small dictionaries are less useful.

Future work. The excellent correction accuracy reached with the perfect dictionary and the number of false friend errors that occur when using the crawled dictionary suggest to replace our naive crawling method with more sophisticated strategies. This is a wide field for future research. We might, for example, measure the similarity between C^{ocr} and a given web page before adding it to the repository, using some IR-based similarity measure. Many other strategies might also help to delimit the number of useless words. As to the number of spelling errors that are found in web pages there are significant differences. The majority of all pages seems to contain a neglectable number of spelling errors. However, a small number of pages was found with an unacceptable number of errors. In the future we plan to identify such pages and to exclude them from the crawl, using dictionaries of spelling errors.

References

- [1] A. Dengel, R. Hoch, F. Hönes, T. Jäger, M. Malburg, and A. Weigel. Techniques for improving OCR results. In H. Bunke and P. Wang, editors, *Handbook of Character Recognition and Document Image Analysis*. World Scientific, 1997.
- [2] W. Francis and H. Kučera. Manual of information to accompany a standard corpus of present-day edited american english, for use with digital computers, 1964. (revised 1971 and 1970) Providence, R.I.: Dept. of Ling., Brown University.
- [3] T. Ho, J. Hull, and S. Srihari. A word shape analysis approach to lexicon based word recognition. *Pattern Rec. Letters*, 1992.
- [4] R. Hoch and T. Kieninger. On virtual partitioning of large dictionaries for contextual post-processing to improve character recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 10(4):273–289, 1996.
- [5] K. Kukich. Techniques for automatically correcting words in texts. *ACM Computing Surveys*, pages 377–439, 1992.
- [6] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.*, 1966.
- [7] K. Taghva and E. Stofsky. OCRSpell: an interactive spelling correction system for OCR errors in text. *International Journal of Document Analysis and Recognition*, 3:125–137, 2001.
- [8] F. Weigel, S. Baumann, and J. Rohrschneider. Lexical post-processing by heuristic search and automatic determination of the edit costs. In *Proc. of the Third International Conference on Document Analysis and Recognition (ICDAR 95)*, pages 857–860, 1995.

Topic \ Dict.	D_{\downarrow}^{EG+}	D^{EG+}	D^{EW}	D_{EG+}^{EW}	D^{perf}
Botany (E)	OCR-accuracy (normal tokens): 97.09				
lex. cov.	80.93	88.74	97.09	97.14	100
corr. acc.	97.09	97.09	97.51	97.54	98.92
false fr.	13	16	20	20	0
no chance	70	43	9	6	0
insp. rate	19.73	12.59	4.90	4.80	3.15
Fishes (E)	OCR-accuracy (normal tokens): 98.95				
lex. cov.	98.07	98.55	99.10	99.40	100
corr. acc.	99.01	99.00	99.08	99.06	99.45
false fr.	30	30	36	36	10
no chance	10	10	7	7	8
insp. rate	2.48	2.02	1.35	1.10	1.02
Holocaust (E)	OCR-accuracy (normal tokens): 98.97				
lex. cov.	96.27	97.83	99.19	99.37	100
corr. acc.	99.01	98.97	99.10	99.07	99.66
false fr.	16	20	23	27	0
no chance	7	4	1	1	3
insp. rate	4.38	2.77	1.38	1.11	1.14
Rom. Emp. (E)	OCR-accuracy (normal tokens): 98.79				
lex. cov.	92.70	97.64	99.45	99.48	100
corr. acc.	98.83	98.83	99.07	99.06	99.67
false fr.	20	22	38	39	6
no chance	21	19	3	3	0
insp. rate	3.97	2.96	1.16	1.12	1.38
Mushrooms (E)	OCR-accuracy (normal tokens): 98.95				
lex. cov.	98.52	98.94	99.64	99.68	100
corr. acc.	99.25	99.25	99.34	99.39	99.68
false fr.	12	14	23	24	0
no chance	5	4	2	1	4
insp. rate	2.08	1.83	0.95	0.91	1.13
Neurology (E)	OCR-accuracy (normal tokens): 98.81				
lex. cov.	98.31	99.06	99.83	99.83	100
corr. acc.	99.06	99.03	99.51	99.42	99.87
false fr.	13	17	19	23	0
no chance	5	2	1	1	1
insp. rate	2.59	1.84	1.05	0.99	1.22

Table 3. Results for english corpora.

Topic \ Dict.	D_{\downarrow}^{GE+}	D^{GE+}	D^{GW}	D_{GE+}^{GW}	D^{perf}
Botany (G)	OCR-accuracy (normal tokens): 95.45				
lex. cov.	80.06	90.12	91.32	94.47	100
corr. acc.	95.49	95.53	95.59	95.78	97.98
false fr.	40	55	53	69	11
no chance	93	57	46	39	14
insp. rate	21.19	11.50	10.13	7.18	5.07
Fishes (G)	OCR-accuracy (normal tokens): 98.43				
lex. cov.	77.56	89.72	92.28	93.24	100
corr. acc.	98.43	98.43	98.45	98.46	99.65
false fr.	22	30	35	37	3
no chance	51	29	27	24	2
insp. rate	22.08	9.99	7.64	6.59	2.03
Holocaust (G)	OCR-accuracy (normal tokens): 97.92				
lex. cov.	91.83	96.62	97.85	98.42	100
corr. acc.	98.03	98.26	98.53	98.69	99.47
false fr.	15	20	20	24	2
no chance	29	13	15	11	2
insp. rate	8.35	4.37	3.39	2.86	2.52
Rom. Emp. (G)	OCR-accuracy (normal tokens): 98.55				
lex. cov.	84.29	90.92	96.51	96.89	100
corr. acc.	98.55	98.56	98.59	98.58	99.62
false fr.	14	22	20	26	4
no chance	43	29	19	13	0
insp. rate	15.24	9.07	4.00	3.52	1.45
Mushrooms (G)	OCR-accuracy (normal tokens): 97.45				
lex. cov.	77.66	86.09	92.11	93.07	100
corr. acc.	97.45	97.47	97.55	97.51	99.14
false fr.	26	30	30	33	14
no chance	64	52	38	31	3
insp. rate	21.80	13.73	7.99	7.03	2.36
Neurology (G)	OCR-accuracy (normal tokens): 97.32				
lex. cov.	81.48	90.33	94.53	95.43	100
corr. acc.	97.42	97.38	97.70	97.70	99.47
false fr.	28	35	38	43	4
no chance	56	38	20	17	6
insp. rate	18.43	10.18	6.38	5.39	3.05

Table 4. Results for german corpora.