# Using SMT for OCR Error Correction of Historical Texts

**Haithem Afli, Zhengwei Qiu, Andy Way and Páraic Sheridan**

ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
{hafli, zqui, away, psheridan}@computing.dcu.ie

**Abstract**

A trend to digitize historical paper-based archives has emerged in recent years, with the advent of digital optical scanners. A lot of paper-based books, textbooks, magazines, articles, and documents are being transformed into electronic versions that can be manipulated by a computer. For this purpose, Optical Character Recognition (OCR) systems have been developed to transform scanned digital text into editable computer text. However, different kinds of errors in the OCR system output text can be found, but Automatic Error Correction tools can help in performing the quality of electronic texts by cleaning and removing noises. In this paper, we perform a qualitative and quantitative comparison of several error-correction techniques for historical French documents. Experimentation shows that our Machine Translation for Error Correction method is superior to other Language Modelling correction techniques, with nearly 13% relative improvement compared to the initial baseline.

## 1. Introduction

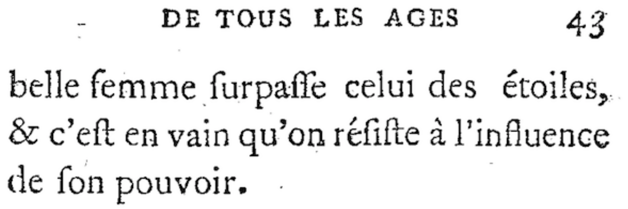
Optical Character Recognition (OCR) systems transform a document image into character-code text. In this process, a document image is segmented into character images in the linear reading order using image-analysis heuristics. An automatic classifier is then applied to determine the character code that most likely corresponds to each character image. However, such systems are still imperfect. There are often mistakes in the scanned texts as the OCR system occasionally misrecognizes letters and falsely identifies scanned text, leading to misspellings and linguistic errors in the output text (Niklas, 2010). In this context, it is appropriate to investigate how we can automatically correct such OCR outputs. This paper explores the effectiveness of OCR output error correction. The correction uses an improved noisy channel model, language modelling and statistical machine translation (SMT) to correct OCR errors. The remainder of the paper is organized as follows: Section 2 provides background information on OCR errors; Section 3 presents related work and our error-correction methodology; Section 4 discusses the results; Section 5 concludes and provides directions for further investigation.

## 2. OCR errors

Optical Character Recognition of historical French documents presents several challenges, including:

- Old-style language: some old letters can be used in these kinds of texts such as the *"long s"* in old French writing which is often confused with the lower-cased *"f "* in various Roman typefaces and in *blackletter*. In the example of Figure 1, *"belle femme surpasse"* can be recognised by automatic systems as *"belle femme furpaffe"*.

- Punctuation errors: especially when we have a poor scanning quality (like in Figure 2), punctuation character misrecognition can cause commas or full stops to often occur in the wrong positions.



DE TOUS LES AGES 43

belle femme ſurpaſſe celui des étoiles,
& c'eſt en vain qu'on réſiſte à l'influence
de ſon pouvoir.

Figure 1: Example of a French sentence in old-style language using a *"long s"* character.

- Character format: variations in font can also prevent accurate character recognition which leads to wrong word recognition (*e.g.*"*èn*" instead of "*en*" ). It can also cause the case-sensitivity errors when lower and upper case characters can be mixed up.

- Character Insertion, Deletion and Substitution (IDS): it is often the case that one or more characters are substituted or deleted, or that a character is wrongly inserted in the middle of a word.

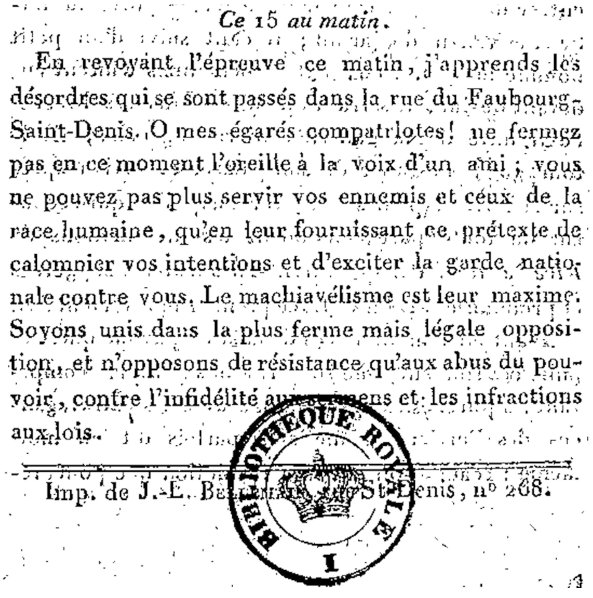
- Segmentation errors: different spacings in lines, words or characters lead to misrecognitions of white-spaces in some cases which can cause segmentation errors (*e.g.*"*bellefemme*" instead of "*belle femme*").

- Word meaning: some misrecognized characters can generate new words which are often wrong in context but correctly spelled (*e.g.*"*red dear*" instead of "*red deer*").

## 3. OCR error correction

### 3.1. Related work

Much research has been done on OCR error corrections, with different strategies including the improvement of visual and linguistic techniques and the combination of several OCR system outputs (Hong, 1995; Schäfer and Weitz,



*Ce* 15 *au matin.*

En revoyant l'épreuve ce matin, j'apprends les désordres qui se sont passés dans la rue du Faubourg-Saint-Denis. O mes égarés compatriotes! ne fermez pas en ce moment l'oreille à la voix d'un ami; vous ne pouvez pas plus servir vos ennemis et ceux de la race humaine, qu'en leur fournissant ce prétexte de calomnier vos intentions et d'exciter la garde nationale contre vous. Le machiavélisme est leur maxime. Soyons unis dans la plus ferme mais légale opposition, et n'opposons de résistance qu'aux abus du pouvoir, contre l'infidélité [illegible] et les infractions aux lois.

Imp. de J.-L. B[illegible] St-Denis, n° 268.

Figure 2: Example of French historical document from National Library of France.

2012).
Post-OCR correction, which represents the basis of correction in this paper, is one of the main directions in this field. In this way, we can consider the OCR system as a black-box, since this technique does not rely on any parameters specific to the OCR system.
The goal of post-processing is to detect and correct errors in the OCR output after the input image has been scanned and completely processed. The obvious way to correct OCR misspellings is to edit the output text manually using translators or linguists. This method requires a continuous manual human intervention which is to a costly and time-consuming practice. There are two main existing approaches to automatically correct OCR outputs. The first approach is based on lexical error correction (Niwa et al., 1992; Hong, 1995; Bassil and Alwani, 2012). In this method, a lexicon is used to spell-check OCR-recognized words and correct them if they are not present in the dictionary. Although this technique is easy to implement and use, it still has various limitations that prevent it from being the perfect solution for OCR error correction (Hong, 1995). It requires a wide-ranging dictionary that covers every word in the language. Existing linguistic resources can usually target a single specific language in a given period, but cannot, therefore, support historical documents.

The second type of approach in OCR post-processing is context-based error correction. Those techniques are founded on statistical language modelling and word $n$-grams. It aims at calculating the likelihood that a word sequence appears Tillenius (1996); Magdy and Darwish (2006). Applying this technique on historical documents might be challenging because work on building corpora for this kind of task (on old-style languages) has been very limited. Furthermore, when many consecutive corrupted words are encountered in a sentence, it is difficult to choose good candidate words. This is illustrated in Figure 2 where OCR recognised *",,Ea ffSYftÿânt,.l'éprçuvè .ce matih, j'app.rends jos dé§Qt;d4es.qui.se Sontpasses dansja, rne du F.auô.uçg- "* instead of *"En revoyant l'épreuve ce matin, j'apprends les désordres qui se sont passés dans la rue du Faubourg"* because of the very poor quality of the document.
In this paper we are using a corpus on old-style French OCR-ed data from the $17^{th}$, $18^{th}$ and $19^{th}$ centuries in order to test this kind of approach with new OCR error correction techniques.

### 3.2. Statistical Machine Translation method

The idea centres on using an SMT system trained on OCR output texts which have been post-edited and manually corrected. SMT systems handle the translation process as the transformation of a sequence of symbols in a source language into another sequence of symbols in a target language. Generally the symbols dealt with are the words in two languages. We consider that our SMT system will translate OCR output to corrected text in the same language following the work of Fancellu et al. (2014); Afli et al. (2015). In fact, using the standard approach of SMT we are given a sentence (sequence of OCR output words) $s^M = s_1...s_M$ of size $M$ which is to be translated into a corrected sentence $t^N = t_1...t_N$ of size $N$ in the same language (French in our case). The statistical approach aims at determining the translation $t^*$ which maximize the posterior probability given the source sentence. Formally, by using the Bayes'rule, the fundamental equation is (1):

$$t^* = \arg\max_t Pr(t|s) = \arg\max_t Pr(s|t)Pr(t) \qquad (1)$$

It can be decomposed, as in the original work of (Brown et al., 1993), into a language model probability $Pr(t)$, and a translation model probability $Pr(s|t)$. The language model is trained on a large quantity of correct French data, and the translation model is trained using a bilingual text aligned at sentence (segment) level, $i.e.$ an OCR output for a segment and its ground-truth obtained manually, so-called bitexts. As in most current state-of-the-art systems, the translation probability is modelled using the log-linear model (Och and Ney, 2002) in (2):

$$P(t|s) = \sum_{i=0}^{N} \lambda_i h_i(s,t) \qquad (2)$$

where $h_i(s,t)$ is the i$^{th}$ feature function and $\lambda_i$ its weight (determined by an optimization process). We call this method *"SMT_cor "* in the rest of this paper. We used only Word-based model for this task because Afli et al. (2015) demonstrated in a similar work that it improves the results better than the character-based model. In the special case of OCR correction, the source and target languages are the same (French in this case) which should not require the use of a reordering model.

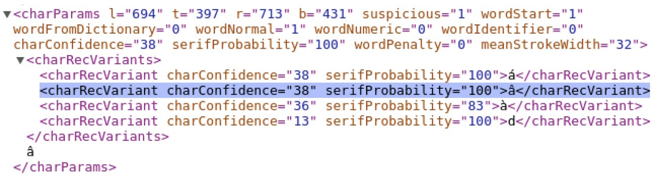
### 3.3. Language Modelling

Language Modelling is a flexible method that can be used to calculate the likelihood that a word sequence would appear. Using this technique, the candidate correction of an error might be successfully found using the "Noisy Channel Model" Mays et al. (1991). Considering the sentence *"I saw a red daer"*, the error-correction system would replace the word *"daer"* by *"dear"* or *"deer"*, and then a language model will likely indicate that the word bigram *"red deer"* is much more likely the the bigram *"red dear"*. For each OCR-ed word $w$ we are looking for the word $c$ that is the most likely spelling correction for that word (this "correction" may be the original word itself). Following the standard methodology we are trying to find the correction $c$ that maximizes $P(c|w)$, as in (3):

$$\arg\max_c Pr(w|c)Pr(c) \tag{3}$$

Here $P(c)$ is the probability that $c$ is the intended word. This is called the *language model*. $P(w|c)$ is the probability that the OCR system recognises $w$ when $c$ is intended. This is called the *error model* or the *noisy channel model*. We call this method *"LM_cor"* in the rest of this paper.

Some OCR systems like *ABBYY*[1] have an internal evaluation that can detect a suspicious character in the word $w$ and provide some possible candidates as in Figure 3.



```
▼<charParams l="694" t="397" r="713" b="431" suspicious="1" wordStart="1"
 wordFromDictionary="0" wordNormal="1" wordNumeric="0" wordIdentifier="0"
 charConfidence="38" serifProbability="100" wordPenalty="0" meanStrokeWidth="32">
  ▼<charRecVariants>
     <charRecVariant charConfidence="38" serifProbability="100">á</charRecVariant>
     <charRecVariant charConfidence="38" serifProbability="100">â</charRecVariant>
     <charRecVariant charConfidence="36" serifProbability="83">à</charRecVariant>
     <charRecVariant charConfidence="13" serifProbability="100">d</charRecVariant>
   </charRecVariants>
   â
 </charParams>
```

Figure 3: Example of a suspicious character and its proposed candidates. The OCR system proposes 4 candidates ("á","â","à" and "d") and after the calculation of an internal character confidence the second candidate "â" is chosen.

This internal evaluation can reduce the problem of estimating the *error model* in the previous method by using the language model only on suspicious words (that contain suspicious characters). In what follows, we call this method *"Sus_cor"*.

# 4. Experimental Results

## 4.1. Data description

For the training of our models, we used a corpus of nearly 60 million OCR output words obtained from scanned documents, developed by Afli et al. (2015). The original data composes about 90 million words, but we only used texts from $19^{th}$, $18^{th}$ and $17^{th}$ centuries. We used the corrected part of this corpus to develop language model. Next, the OCR output sentences and the manually corrected version were aligned at word level and this bitext was used for our SMT error correction method. For testing, we processed 10 documents in old-style French from the $17^{th}$ century using the *ABBYY* OCR tool and manually corrected the text output in order to provide a correction reference for the evaluation. The statistics of all corpora used in our experiments can be seen in Table 1.

| bitexts | # OCR tokens | # ref tokens |
|---|---|---|
| smt_17 | 1.98 M | 1.96 M |
| smt_18 | 33.49 M | 33.40 M |
| smt_19 | 23.08 M | 22.9 M |
| dev17 | 8638 | 8638 |
| test17 | 1660 | 1660 |

Table 1: Statistics of MT training, development and test data available to build our systems.

## 4.2. Testing the Models

For all of the different techniques used in this paper, the language model was built using the KenLM[2] toolkit (Heafield, 2011) with Kneser-Ney smoothing (Kneser and Ney, 1995) and default backoff. For the $SMT\_cor$ method, an SMT system is trained on all available parallel data. Our SMT system is a phrase-based system (Koehn et al., 2003) based on the Moses SMT toolkit (Koehn et al., 2007). It is constructed as follows. First, word alignments in both directions are calculated, using GIZA++ (Gao and Vogel, 2008). Phrases are extracted using the default settings in the Moses toolkit. The parameters of our system were tuned on a development corpus, using Minimum Error Rate Training (Och, 2003).

We combined our different systems in order to test the impact of successive corrections using different models. Our different techniques do not always correct the same errors, so a combination can be beneficial. Two paths were pursued to explore the effect of the combination of the $SMT\_cor$ and $LM\_cor$ methods. In the first one, we begin by generating the corrected output of the $LM\_cor$ which is then processed by the $SMT\_cor$ system. We call this combination $LM\_cor + SMT\_cor$. In the second one, we reverse the use of the two systems and we call this combination $SMT\_cor + LM\_cor$.

## 4.3. Results

For the evaluation, we used Word Error Rate (WER) which is derived from Levenshtein distance Levenshtein (1966). We compare results on test data of all different methods used in our experiments to the baseline results which represent scores between the OCR output and the corrected reference (which we call *OCR-Baseline*). Table 2 reports on the percentage of Correctness, Accuracy and WER of different system outputs. The best model, using the *SMT_cor* system, was able to decrease nearly 3% of the OCR word errors with 12.99% relative improvement.

It can also be observed that the LM_cor and the Sus_cor systems did not improve the results in this task of correcting errors in OCR-ed historical documents. The failure to find a proper correction using these two systems was generally due to words from the $17^{th}$ century that were changed in the $18^{th}$ or $19^{th}$ centuries. As our training data are almost are from $18^{th}$ and $19^{th}$ centuries, the language modelling methodology can even introduce more errors when correcting test data from $17^{th}$ century. This observation is confirmed by the combination of the

[1] http://www.abbyy.com/

[2] https://kheafield.com/code/kenlm

|  |  |  |  | **% Relative** | Errors |  |  |
|---|---|---|---|---|---|---|---|
| Systems | Correctness | **Accuracy** | **WER** | **improvement** | % Insertion | % Deletion | % Substitution |
| OCR-Baseline | 81.14 | 77.59 | 22.41 | - - | 15.86 | 12.90 | 71.23 |
| LM_cor | 79.22 | 75.66 | 24.34 | - 8.61 | 14.60 | 11.88 | 73.51 |
| Sus_cor | 79.76 | 77.05 | 22.95 | - 2.41 | 11.81 | 12.33 | 75.85 |
| SMT_cor | **83.92** | **80.48** | **19.52** | **+ 12.99** | 17.59 | 15.43 | 66.97 |
| Systems Combination |  |  |  |  |  |  |  |
| LM_cor + SMT_cor | 81.45 | 78.01 | 21.99 | + 1.87 | 15.61 | 13.69 | 70.68 |
| SMT_cor + LM_cor | 81.87 | 78.43 | 21.57 | + 3.74 | 15.92 | 13.96 | 78.43 |
| Sus_cor+ SMT_cor | 82.65 | 79.94 | 20.06 | + 10.48 | 13.51 | 12.91 | 73.57 |

Table 2: Percentage of Relative improvement, WER, Accuracy and Correctness results of our System Correction on test17.

best model (*SMT_cor*) with the *LM_cor* and the *Sus_cor* systems, where we can see that errors introduced by the language-modelling method decreases the overall result of the *SMT_cor* correction. We anticipate that more training data in the same style of the test data would improve the correction of all proposed methods in this paper and especially the language modelling methodology.

## 5. Conclusion

In this paper, we have investigated the use of different post-OCR correction methods to correct errors in OCR-ed historical documents. The results show the superiority of the SMT method in the case where training data in the same style of the test data is lacking, compared to the different language modelling methods used. Our best model outperforms the OCR-Baseline by nearly 13% relative improvement. Accordingly we believe that our methods can be a good way to resolve the problem of correcting OCR errors for historical texts. We plan to test these methods on other different languages and types of data.

## 6. Acknowledgments

This research is supported by Science Foundation Ireland through the TIDA Programme (Grant 14/TIDA/2384) in the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University.

## References

Haithem Afli, Loïc Barrault, and Holger Schwenk. OCR Error Correction Using Statistical Machine Translation. *16th International Conference on Intelligent Text Processing and Computational Linguistics*, Cairo, Egypt 2015.

Youssef Bassil and Mohammad Alwani. OCR Post-Processing Error Correction Algorithm Using Google's Online Spelling Suggestion. *Journal of Emerging Trends in Computing and Information Sciences*, 3:90–99, 2012.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311, 1993.

Federico Fancellu, Andy Way, and Morgan O'Brien. Standard language variety conversion for content localisation via SMT. *17th Annual Conference of the European Association for Machine Translation*, pages 143–149, Dubrovnik, Croatia 2014.

Q. Gao and S. Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, in Columbus, Ohio, USA 2008.

Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, UK, 2011.

Tao Hong. *Degraded Text Recognition Using Visual and Linguistic Context.* PhD thesis, University of New York, NY, USA 1995.

Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 181–184, Detroit, Michigan, USA, 1995.

P. Koehn, Franz J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 48–54, Edmonton, Canada 2003.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic 2007.

V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

Walid Magdy and Kareem Darwish. Arabic OCR Error Correction Using Character Segment Correction, Language Modeling, and Shallow Morphology. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 408–414, Sydney, Australia 2006.

Eric Mays, Fred J. Damerau, and Robert L. Mercer. Context based spelling correction. *Information Processing and Management*, 27(5):517–522, 1991.

Kai Niklas. *Unsupervised Post-Correction of OCR Errors*. PhD thesis, Leibniz University Hannover, in Germany 2010.

Hisao Niwa, Kazuhiro Kayashima, and Yasuham Shimeki. Postprocessing for character recognition using keyword information. In *IAPR Workshop on Machine Vision Applications*, volume MVA'92, pages 519–522, Tokyo, Japan 1992.

Franz J. Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 160–167, Sapporo, Japan 2003.

Franz Josef Och and Hermann Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, USA, 2002.

Ulrich Schäfer and Benjamin Weitz. Combining OCR outputs for logical document structure markup: Technical background to the ACL 2012 contributed task. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 104–109, Jeju Island, Korea 2012.

Mikael Tillenius. Efficient generation and ranking of spelling error corrections. Technical report, Royal Institute of Technology, Stockholm, Sweden 1996.