Integrating Optical Character Recognition and Machine Translation of Historical Documents

Haithem Afli and Andy Way
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
{haithem.afli, andy.way}@adaptcentre.ie

Abstract

Machine Translation (MT) plays a critical role in expanding capacity in the translation industry. However, many valuable documents, including digital documents, are encoded in non-accessible formats for machine processing (*e.g.*, Historical or Legal documents). Such documents must be passed through a process of Optical Character Recognition (OCR) to render the text suitable for MT. No matter how good the OCR is, this process introduces recognition errors, which often renders MT ineffective. In this paper, we propose a new OCR to MT framework based on adding a new OCR error correction module to enhance the overall quality of translation. Experimentation shows that our new system correction based on the combination of Language Modeling and Translation methods outperforms the baseline system by nearly 30% relative improvement.

1 Introduction

While research on improving Optical Character Recognition (OCR) algorithms is ongoing, our assessment is that Machine Translation (MT) will continue to produce unacceptable translation errors (or non-translations) based solely on the automatic output of OCR systems. The problem comes from the fact that current OCR and Machine Translation systems are commercially distinct and separate technologies. 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). Works involved in improving translation services purchase off-the-shelf OCR technology but have limited capability to adapt the OCR processing to improve overall machine translation performance. In this context, it is appropriate to investigate the integration of OCR and MT for improved translation accuracy. A novel integration of OCR, error correction and MT technology that results in overall improvements in translation output quality to the point of acceptance by professional translators for post-editing, would have a profound effect on the economics of translation in high-value (expensive) domains such as Historical documents translation. This paper explores the effectiveness of OCR output error correction and its impact on automatic translation. The correction uses a combination of language modelling and statistical machine translation (SMT) methods to correct OCR errors. Our goal is to address the question of whether a shared and novel integration of language processing encompassing OCR, error correction and MT could significantly improve the final translation quality of the text which has initially been OCRed. The remainder of the paper is organized as follows: Section 2 presents related work on OCR error correction; Sections 3 and 4 describe the proposed OCR to MT framework and the feasibility experiments conducted. Section 5 reports and discusses about the results and the directions for further work are provided in Section 6.

2 OCR error correction

2.1 Related work

The current state of the OCR output translation includes expensive manual intervention in order to correct the errors introduced in processing texts through OCR, or simply takes the approach of undertaking manual re-creation of the document in a machine-processable form. Alternatively, the document is retained in its original format and is provided as an 'image' to a professional translator to translate into the target language, though this means that the original 'source' language is not available for inclusion in Translation Memory for future use. In the worst case, the text is manually processed in its source text and then professionally translated into the target text without any automated processing at all.

A lot of research has been carried out on OCR error correction, with different strategies including the improvement of visual and linguistic techniques as well as combining several OCR system outputs (Hong, 1995; Schäfer and Weitz, 2012; Springmann and Ldeling, 2016). Such post-OCR correction, which represents the focus of this paper, is one of the main directions in this domain. 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 the OCR misspellings is to edit the output text manually using translators or linguists. This method requires continuous manual human intervention which is a costly and time-consuming practice. There are two main existing approaches to automatically correct the 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, 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. These techniques are founded on statistical language modelling and word n-grams, and aims to calculate the likelihood that a particular word sequence appears (Tillenius, 1996; Magdy and Darwish, 2006). Applying this technique on historical documents is challenging because the works on building corpora for this kind of task has been very limited. Furthermore, when many consecutive corrupted words are encountered in a sentence, it is difficult to choose the good candidate words. In this paper we conducted our experiments using a corpus of old-style French OCR-ed data from the 17^{th}, 18^{th} and 19^{th} centuries in order to verify the applicability of our new OCR-to-MT framework.

2.2 Translation method

This technique centres on using an SMT system trained on the 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 the 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., 2016).

In fact, using the standard approach of SMT we are given a sentence (a 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 maximizes the posterior probability given the source sentence. Formally, by using Bayes' rule, the fundamental equation is (1):

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

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 French texts 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. As in most current state-of-the-art systems, the translation probability is modelled using the log-linear model in (2):

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

where $h_i(s,t)$ is the i$^t h$ feature function and λ_i its weight (determined by an optimization process). We call this method *"SMT_cor "* in the rest of this paper. As (Nakov and Tiedemann, 2012; Tiedemann and Nakov, 2013) demonstrated, closely related languages largely overlap in vocabulary and have a strong syntactic and lexical similarities. We assume that we do not need to use the reordering model in the task of error correction in the same language.

2.3 Language Modelling

Language Modelling is the field of creating models for writing text so that we can assign a probability to a sequence of n consecutive words. 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 drink a baer*', the error correction system would identify '*bear*' or '*beer*' as possible replacements for the non-word '*baer*', and then a language model would most likely indicate that the word trigram '*drink a beer*' is much more likely than '*drink a bear*'. Accordingly, for each OCR-ed word w we are looking for the word c that is the most likely spelling correction for that word (which may indeed be the original word itself).

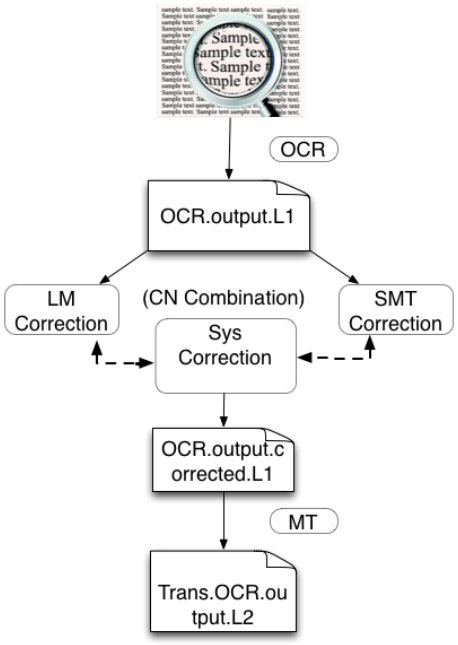


Figure 1: The proposed OCR-to-MT framework.

3 OCR-to-MT framework

The basic system architecture is depicted in Figure 1. We can distinguish three steps: automatic character recognition (OCR), error correction (Sys Correction) and machine translation (MT). The OCR system accepts original documents in language L1 (French in our case) and generates an automatic transcription. This text is then corrected by two different correction systems based on language modelling and SMT

methods described in the previous section. Different correction systems can generate multiple input hypotheses with varying confidence for the combination system based on confusion networks. The final corrected text in L1 forms the input to the MT system. We anticipate that the automatic correction will improve the quality of the final translation to the language L2 (English in our case). Accordingly, this framework sets out to address the question of whether a shared and novel integration of language processing components from both OCR and MT can significantly improve the final translation quality of text which has initially been OCR-ed.

4 Impact of Error Correction on Automatic Translation

The proposed OCR-to-MT framework raises several issues. Each step can introduce a certain number of errors. It is important to highlight the feasibility of the approach and the impact of each module on the final automatic translation. Thus, we conducted three different types of experiments, described in Figure 2.

In the first experiment (*Exp. 1*) we use the OCR reference (*Ref.OCR.fr*) as input to the MT system. This is the most favourable condition, as it simulates the case where the OCR and the Error Correction systems do not commit any error. Accordingly, we consider this as the reference during the automatic evaluation process. In the second experiment (*Exp. 2*) – the baseline experiment – we use the OCR output (*OCR.output.fr*) directly as input to the MT system without any correction. Finally, the third experiment represents the complete proposed framework, described in Section 3.

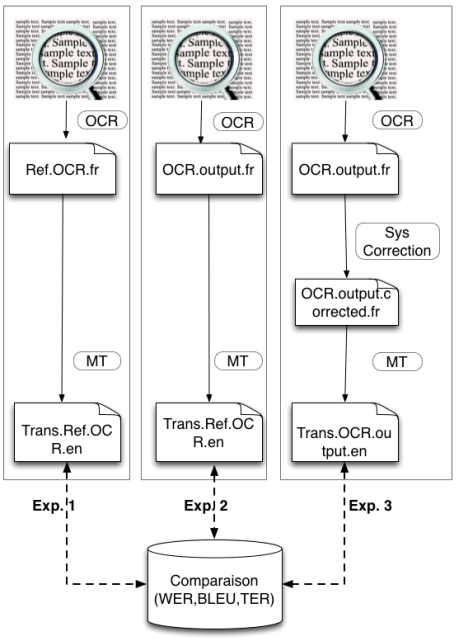


Figure 2: Different experiments to analyze the impact of the Error Correction module.

5 Experimental Results

5.1 Data and systems description

For the training of our models, we used a corpus of nearly 58 million OCR output words obtained from scanned documents, developed by (Afli et al., 2015). We used the corrected part of this corpus for the 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 used OCR-ed French data (dev17) from the 17th century, manually corrected. 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	9013	8946

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

For all of the different techniques used in this paper, the language model was built using the KenLM toolkit with Kneser-Ney smoothing 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). Word alignments in both directions are calculated, using a multi-threaded version of the GIZA++ tool (Gao and Vogel, 2008).

The parameters of our system were tuned on a development corpus, using Minimum Error Rate Training (Och, 2003). We combined our two systems using a Confusion Network (CN) combination system based on the work of (Wu et al., 2012). We call this combination $LM\_cor + SMT\_cor$.

5.2 Results

In order to evaluate the effectiveness of error correction, we used Word Error Rate (WER) which is derived from Levenshtein distance (Levenshtein, 1966). We compare results on the test data of the two different methods used in our experiments and their combination, against the baseline results which represent scores between OCR output and the corrected reference (called *OCR-Baseline*).

Table 2 reports on the percentage of Correctness, Accuracy and WER of different system outputs. The best model, using the *CN Comb.* system was able to decrease 5.39% of the OCR word errors (29.42% relative improvement). It can also be observed that the SMT_cor system improves the results more than LM_cor. Nonetheless, both underperform compared to the *CN Comb.* system. This is due to the fact that the two methods are not always correcting the same errors, so the CN combination can be beneficial in this case.

Systems	Correctness	Accuracy	WER
Baseline	83.92	81.68	18.32
LM_cor	84.82	82.57	17.43
SMT_cor	87.64	86.06	13.94
CN Comb.			
LM_cor + SMT_cor	**89.10**	**87.07**	**12.93**

Table 2: Word Error Rate (WER), Accuracy and Correctness results on on dev17 OCR-corrected data.

For the translation evaluation we used BLEU-4 score (Papineni et al., 2002), Smoothed BLEU (Lin and Och, 2004) and TER (Snover et al., 2006) calculated between the output of *Exp. 1* (our reference) and *Exp. 2* output (the baseline) or *Exp. 3* output (our proposed framework).

Table 3 lists the results of the two translation outputs from *Exp. 1* and *Exp. 2*. It shows that our proposed framework is very capable of correcting the final translation of the OCR-ed documents.

5.3 Analysis and Discussion

In order to better understand the impact of the error correction process and the problems of OCR'ed historical document translation, we prepared a manual human translation of our test set based on the

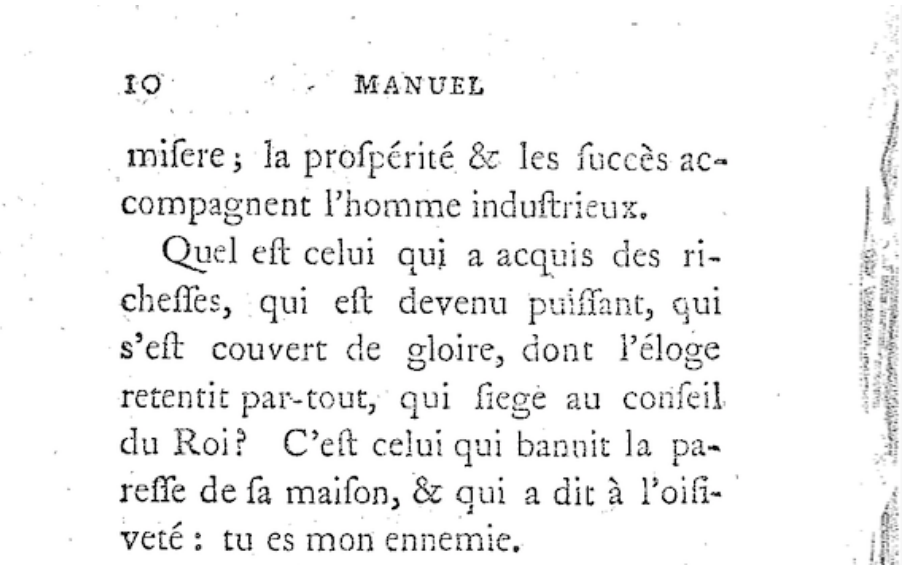
https://kheafield.com/code/kenlm

The source is available at `http://www.cs.cmu.edu/~qing/`

Systems	BLEU-4	Smooth BLEU	TER
Exp. 2	24.53	38.29	57.32
Exp. 3	**69.43**	**70.15**	**21.62**

Table 3: BLEU-4, Smooth BLEU and TER results on dev17 OCR-translated data.

transformation of the old French language to the current one and its translation to current English language without any modification on the original format. We find that comparing to the manual translation, the system can not get the correct context of the documents lines because of their short length. As we can see in the figure 3, the sentence starts with the word '*Quel*' and finish with the word '*Roi?*' is segmented on five lines which can cause a translation context problem for the MT system.


IO MANUEL

miſere ; la proſpérité & les ſuccès accompagnent l'homme induſtrieux.

Quel eſt celui qui a acquis des richeſſes, qui eſt devenu puiſſant, qui s'eſt couvert de gloire, dont l'éloge retentit par-tout, qui ſiege au conſeil du Roi? C'eſt celui qui bannit la pareſſe de ſa maiſon, & qui a dit à l'oiſiveté : tu es mon ennemie.

Figure 3: Exemple of Historical French document with short length of its lines.

As almost all current MT systems are translating document line by line, we can say that the context of the translations is line-based even when we try to adapt the system to the domain of the document. The particularity of historical documents can cause a problem of context translation without a pre-processing of the corrected OCR output. Our results presented in table 4 show that the transformation of the OCR'ed data to one sentence per line as a pre-processing can improve the automatic translation results from 12.59 to 18.92 BLEU points which is a very important improvement.

Manual translation	without pre-processing	sentence per line
BLEU-4	12.59	18.92

Table 4: BLEU-4 results on dev17 OCR-translated data with- and without pre-processing compared to manual translation.

This experiment can open the way of thinking about improving our current MT methods and systems by getting the document-level context of translation.

6 Conclusion

In this paper, we presented a new framework of OCR-ed document translation. The proposed method consists of the integration of a new error correction system prior to the translation phase *per se*. We validate the feasibility of our approach using a set of experiments to analyze the impact of our OCR error correction module on the final translation. Experiments conducted on old-style French data showed that

our methodology improves the quality of the translation of OCR documents. Accordingly, we believe that our method can be a good way to resolve the problem of correcting OCR errors for historical texts. We plan to test it on other different languages and types of data and try to integrate the correction system inside the OCR system architecture itself.

Acknowledgements

This research is supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (`www.adaptcentre.ie`) at Dublin City University. We would like to thank Wided Baazaoui for her useful and knowledgeable help on preparing the development- and test-set of this work.

References

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

Haithem Afli, Zhengwei Qiu, Andy Way, and Páraic Sheridan. 2016. Using SMT for OCR error correction of historical texts. In *Proceedings of LREC-2016*, pages 962–965, Portorož, Slovenia.

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

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

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

Q. Gao and S. Vogel. 2008. 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.

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

P. Koehn, Franz J. Och, and D. Marcu. 2003. 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.

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. 2007. 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.

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

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 605–612, Barcelona, Spain.

Walid Magdy and Kareem Darwish. 2006. 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.

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

Preslav Nakov and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 301–305.

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

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

Franz J. Och. 2003. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Ulrich Schäfer and Benjamin Weitz. 2012. 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.

S. Snover, B. Dorr, R. Schwartz, M. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Uwe Springmann and Anke Ldeling. 2016. OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. *ArXiv e-prints*.

Jörg Tiedemann and Preslav Nakov. 2013. Analyzing the use of character-level translation with sparse and noisy datasets. *Proceedings of Recent Advences in Natural Language Processing*, pages 676–684.

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

Xiaofeng Wu, Tsyoshi Okita, Josef van Genabith, and Qun Liu. 2012. System combination with extra alignment information. In *Second ML4HMT Workshop (COLING 2012)*, pages 37–44, Mumbai, India.