OCR Error Correction Using Statistical Machine Translation

Haithem Afli, Loïc Barrault and Holger Schwenk

Université du Maine,
Avenue Olivier Messiaen F-72085 - LE MANS, France
FirstName.LastName@lium.univ-lemans.fr

Abstract. In this paper, we explore the use of a statistical machine translation system for optical character recognition (OCR) error correction. We investigate the use of word and character-level models to support a translation from OCR system output to correct french text. Our experiments show that character and word based machine translation correction make significant improvements to the quality of the text produced through digitization. We test the approach on historical data provided by the National Library of France. It shows a relative Word Error Rate reduction of 60% at the word-level, and 54% at the character level.

Keywords: OCR error correction, SMT, post-processing.

1 Introduction

With the advent of digital optical scanners, the trend to digitize (historic) paper-based archives has emerged in recent years. A lot of paper-based books, textbooks, magazines, articles, and documents are being transformed into an electronic version that can be manipulated by a computer. For this purpose, optical character recognition (OCR) systems was developed to translate scanned graphical text into editable computer text. However, those systems are still imperfect. There are often mistakes in the scanned texts as OCR system occasionally misrecognizes letters and falsely identifies scanned text, leading to misspellings and linguistics errors in the output text [12]. In this context, it is appropriate to ask the following question: How can we automatically correct the OCR outputs? This paper proposes a post-processing error correction method for detecting and correcting OCR non-word and real-word errors. The proposed algorithm is based on a statistical machine translation (SMT) system used to "translate" an OCR output to a corrected version based on statistical methods. The paper will be organized as follows: Section 2 presents the related work; Section 3 provides the proposed error correction method; Section 4 describes technical details about using word and character level SMT models; Section 5 reports and discuss the results; and section 6 concludes the paper and provides possible future directions.

OCR output	**References**
stir la place	sur la place
articles 6 et **g** de	articles 6 et 9 de
à la **diarge**	à la charge
avec **1 adhésion**	avec l' adhésion
de **là** cuisse	de la cuisse
il **appuyé** la main	il appuye la main
le dos **èt** sur les	le dos et sur les

Table 1. Example of OCR errors identified in the historical data from the French National Library.

2 Background

2.1 OCR errors

OCR systems transform a document image into character-code text. In this process, a document image is segmented into characters images in the proper reading order using image analysis heuristics. Then, an automatic classifier is applied to determine the character code that most likely correspond to each character image. In the French historical OCR, we have several challenges including:

– Non-word error: this kind of errors occurs if the recognized word is not contained in any dictionary (*e.g.*"*diarge*" instead of "*charge*").
– Segmentation error: spacings in different line, word or character lead to misrecognitions of white-spaces in some cases. It can cause segmentation errors (*e.g.*"*mais on*" instead of "*maison*").
– Punctuation errors: punctuation character misrecognition can cause that commas or points occur more often in wrong places.
– Case sensitivity: lower and upper case characters can be mixed up (*e.g.*"*FraNCe*" instead of "*France*").
– Character format: font variation can also prevents an accurate character recognition which cause wrong word recognition (*e.g.*"*èt*" instead of "*et*" or "*6 et g*" instead of "*6 et 9*").
– Word meaning: some misrecognized characters can generate new words which are often wrong in the context but correctly spelled (*e.g.*"*de là cuisse*" instead of "*de la cuisse*").

Some examples of OCR errors can be seen in Table 1.

2.2 OCR error correction

Much research has been done on OCR errors corrections, which can be divided into three main areas :

– improving visual and linguistic technique by using scanned images [5].

– combining several OCR system outputs in order to select the best one [16].
– post-processing the OCR output text to correct it.

Our method is related to the third kind of methods. In this way, we can consider the OCR system as a black-box, since our method does not rely on any parameter specific to the OCR system.

2.3 OCR output correction by post-processing

The goal of post-processing is to detect and correct misspellings in the OCR output after the input image has been scanned and completely processed. The obvious way to correct OCR errors is to edit the output text manually by linguists. This method requires a continuous manual human intervention which is to some degree regarded as a costly and time-consuming practice. There are two main existing approaches to automatically correct OCR outputs. The first approach is based on a lexical error correction [13, 5, 1]. In this method, a lexicon is used to spell check OCR recognized words and correct them if they are not present on the dictionary. Although this technique is easy to implement and use, it still have various limitations that prevent it to being the perfect solution for OCR error correction [5]. The first one is that it requires a wide-ranging dictionary that covers every single word in the language. However, many morphologically complex languages, such as Arabic, German and Finnish, have enormous numbers of possible words. Another limitation is that conventional dictionary do not support names of regions, geographical locations, some technical keyword and domain specific terms. They normally target a single specific language in a given period, and thus, cannot support historical documents with different styles of writing. The second type of approach in OCR post-processing is the context-based error correction. Those techniques are founded on statistical language modeling and word n-grams. It aims at calculating the likelihood of a word sequence to appear [20, 10]. Considering the sentence *"celte dénégation de l'appelant"*, the error correction system would consider word candidates for replacing the word *"celte"*, *i.e.* *"cette"* or *"celle"*, then a language model would be used to keep the more probable hypothesis (*"cette"* in this case). However, some words are more likely corrected than others because they are more frequent like *e.g.*the stop words, which can result in erroneous corrections. Also when many consecutive corrupted words are encountered in a sentence then it is difficult to consider good candidate words. This is illustrated in Figure 1. OCR recognized

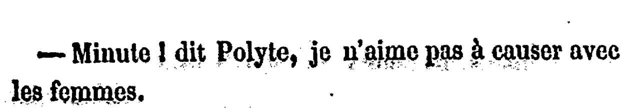
— Minute ! dit Polyte, je n'aime pas à causer avec les femmes.

Fig. 1. Example of a badly recognized sentence.

"Minute 1 dit Polyte, je $ ^}BQ Sf 8 £ W$er avec les femmes." instead of *"Minute ! dit Polyte , je n' aime pas à causer avec les femmes ."* Note that in this case, the OCR output has been processed in order to recompose the full sentence.

3 Statistical Machine Translation method for OCR error correction

3.1 Basic Idea

This papers proposes a new post-processing method for OCR error correction using Statistical Machine Translation technique. The idea centers on using an SMT system trained on OCR output texts post-edited and manually corrected. SMT system 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. In fact, using the standard approach of statistical machine translation 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 equation of statistical machine translation is the following:

$$t^* = \arg\max_t Pr(t|s) = \arg\max_t Pr(s|t)Pr(t) \tag{1}$$

It can be decomposed, as in the original work of [2], 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 most of current state-of-art systems, the translation probability is modeled using the log-linear model, defined as follows:

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

where $h_i(s,t)$ is the i$^t h$ feature function and λ_i its weight (determined by an optimization process). As [11, 19] demonstrated, closely related languages largely overlap in vocabulary and have a strong syntactic and lexical similarities. These similarities motivated them to use a character-level SMT model. We propose to use this method as an additional experiments of word-level SMT models, as we translate two sub-languages with strong similarities. One difficulty with character-level SMT models is their inability to model long-distance word orderings [18]. 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. This assumption will be verified experimentally. Another issue, is that training this kind of model requires that an alignment of the characters in the bitexts is available [21]. Some alignment algorithms specialized in character to phoneme translation have been developed [3, 6]. But [11] has demonstrated that the standard approach based on IBM model 4 alignment [2] is also effective in the case of character-based data. We followed the same method.

3.2 Baseline SMT system

Our SMT system is a phrase-based system [8] based on the Moses SMT toolkit [7]. The 14 standard feature functions are used, namely phrase and lexical translation probabilities in both directions, a word and a phrase penalty and a target language model. It is constructed as follows. First, word alignments in both directions are calculated. We used the multi-threaded version of the GIZA++ tool [4]. Phrases are extracted using the default settings of the Moses toolkit. The parameters of our system were tuned on a development corpus, using Minimum Error Rate Training (MERT) [14]. The language model is trained with the SRILM toolkit [17], on all the available French data. For the evaluation, we used Word Error Rate (WER) metric which is derived from the Levenshtein distance [9] and the BLEU score [15] which is a common metric in the machine translation field. We compare results on development and test data using word-level and character-level SMT systems, and the baseline results which represent scores between OCR output and the corrected reference. We further call it *OCR-Base*.

4 Experiments

For our experiments, we used historical data provided by the National Library of France.

4.1 Data description

To train our models, 90 millions OCR output words were obtained from the scanned documents. The segments were then manually corrected by up to three annotators (the third being called for the segments where the first two disagreed). Next, the OCR output sentences and the manually corrected version were aligned at word level. The development and test corpus was randomly extracted and excluded from the training data of the SMT system. In our experiments on character-level SMT, the characters are considered as words. The special word *SPACE* is used for the original space character and *EOS* for the end of each line. An example of this character segmentation can be seen in the Table 2. Statistics of all corpus used in our experiments can be seen in Table 3 and Table 4 respectively for word- and character-level experiments. In addition , the

OCR output word-level	et , côïnirie ce soût les vôtres . Ils seront encore mieux exécutés . »
Reference word-level	et , comme ce sont les vôtres , ils seront encore mieux exécutés . »
OCR output char-level	e t # , # c ô ï n i r i e # c e # s o û t # l e s # v ô t r e s # . # I l s # s e r o n t # e n c o r e # m i e u x # e x é c u t é s # . # »EOS
Reference char-level	e t # , # c o m m e # c e # s o n t # l e s # v ô t r e s # , # i l s # s e r o n t # e n c o r e # m i e u x # e x é c u t é s # . # »EOS

Table 2. Examples of character segmentation of OCR output and their reference sentences. Special characters # represent a space in the orignal sentences, and EOS differentiate the end-of-sentence character from the others.

data is classified by century from the 18th century (*train_17_BNF*) to the 21th century (*train_20_BNF*). Data with undefined date are labeled *train_0_BNF*. The aim of this classification is to adapt the models to those centuries.

bitexts	# OCR tokens	# ref tokens
train_17_BNF_W	3M	2.9M
train_18_BNF_W	49.4M	49M
train_19_BNF_W	32.5M	32.3M
train_20_BNF_W	300k	300k
train_0_BNF_W	5.2M	5.2M
dev_BNF_W	107.4k	106.5k
tst_BNF_W	112.2k	111.4k

Table 3. Statistics of MT training, development and test data available to build the word-level SMT system.

4.2 Word level SMT system

First of all, an SMT system is trained on all available parallel data. This system uses a 4-gram interpolated back-off language model trained on all corrected french data of the corpus.

On the use of a reordering model As mention in Section 3.1, one question which prize is the following: is the reordering model useful to the SMT system when correcting OCR outputs? In order to verify this assumption, we built two systems which only difference lies in the use of a reordering model. The results are presented in Table 5. We can notice that very similar results were obtained.

bitexts	# OCR tokens	# ref tokens
train_17_BNF_C	15.2M	15.1M
train_18_BNF_C	252.8M	252.2M
train_19_BNF_C	168.8M	168.6M
train_20_BNF_C	1.4M	1.4M
train_0_BNF_C	26.5M	26.5M
dev_BNF_C	277.1k	276.9k
tst_BNF_C	287.9k	287.7k

Table 4. Statistics of MT training, development and test data available to build the character-level SMT system.

System	Development		Test	
	WER	BLEU	WER	BLEU
OCR-Base	5.1	90.66	4.9	90.65
SMT reord.	2.1	96.36	1.9	96.74
SMT no reord.	2.1	96.34	1.9	96.69

Table 5. Results of OCR error correction when the SMT system uses a reordering model or not.

BLEU	dev_BNF_W	tst_BNF_W
OCR-Base	90.66	90.65
Word-level	96.34	96.69

Table 6. BLEU scores results on development and test data with word-level SMT system.

It seems that the reordering model has no impact on the results. Given this result, and for the sake of simplicity, the reordering model will be disabled for the rest of the experiments.

Word-level results Table 6 reports the BLEU score on the development and test data using the word-level SMT system. The results indicates that the SMT system improves the results over the baseline OCR output by around 6 %BLEU points on the development and test data. This tend to prove that our proposed SMT method can effectively decrease the OCR errors. These results are confirmed by the WER scores presented in Table 7. These tables show that OCR error correction using a word level SMT system can provide a decrease in terms of WER (column *Err*) from 4.9% to 1.9%, which represent a great relative improvement of around 60%. Most of these corrections correspond to the deletion of extraneous inserted words (column *Ins*) and more importantly to the correction of erroneous word (column *Sub*).

Dev	Corr	Sub	Del	Ins	WER
OCR-Base	96.0	3.8	0.2	1.1	5.1
SMT Word	98.3	1.4	0.3	0.4	2.1
Abs. Impr.	+2.3	-2.4	+0.1	-0.7	-3
Rel. Impr.		-63%	+50%	-63%	-59%
Test	Corr	Sub	Del	Ins	WER
OCR-Base	96.1	3.7	0.3	1.0	4.9
SMT Word	98.5	1.3	0.2	0.3	1.9
Abs. Impr.	+2.4	-2.4	-0.1	-0.7	-3
Rel. Impr.		-65%	-33%	-70%	-61%

Table 7. WER scores on dev and test data using the word-level system. Lines *Abs. Imp.* and *Rel. Imp.* are respectively the absolute and relative improvements.

System	%BLEU		CER	
	Dev	Test	Dev	Test
OCR-Base	97.36	97.39	1.4	1.3
SMT S.len 2500	99.07	99.11	0.7	0.6
SMT split	98.70	98.71	1.2	1.2

Table 8. Effect of setting the sentence length to 2500 versus splitting sentences on the results in terms of %BLEU and character error rate (CER).

4.3 Character level systems

In this section, the development of a character-level SMT system to translate from OCR output to corrected French text is described. In this model, sentences become longer as we consider each letter to be a unit by itself. This is problematic since the training complexity of the word alignment models is exponentially proportional to the length of the bilingual segments. When building word based systems, it is standard to filter out segments which length is more than 100 words. This is clearly not feasible for character based systems. In order to overcome this problem, we compared two different systems. The first is trained with the maximum sentence length set to 2500 (value determined empirically). The second system is trained with a bitext where the sentences has been split according to the word alignment, following the strong punctuation marks. The results of those systems are presented in Table 8. We observe that increasing the maximum sentence length provide better results, so we used that approach in the following experiments. Note that the training of the alignments with GIZA++ took 6h35 when the sentences are split and more than 63h for the other configuration. Another effect of considering character level SMT is that the vocabulary size is reduced. 464 units compose the vocabulary, which contains upper- and lower-cased letters, punctuation marks and special symbols for spaces and end of sentences. This is to be compared to the 687k word-level vocabulary. One of the advantages is that it is possible to greatly increase the order of the language model. In our experiments, we considered 9-gram interpolated back-off language models. The results of the character-level systems interns of BLEU and CER can be seen respectively in the Tables 9 and 10. We can observe a %BLEU score

%BLEU	dev_BNF_C	tst_BNF_C
OCR-Base	97.36	97.39
SMT Char	99.07	99.11

Table 9. BLEU scores results on development and test data using the character-level SMT system.

Dev.	Corr	Sub	Del	Ins	CER
OCR-Base	99.0	0.8	0.2	0.4	1.4
SMT Char.	99.5	0.3	0.2	0.2	0.7
Abs. Impr.	+0.5	-0.5	0	-0.2	-0.7
Rel. Impr.		-62%	0%	-33%	-50%
Test	**Corr**	**Sub**	**Del**	**Ins**	**CER**
OCR-Base	99.0	0.8	0.2	0.3	1.3
SMT Char.	99.5	0.3	0.2	0.1	0.6
Abs. Impr.	+0.5	-0.5	0	-0.2	-0.7
Rel. Impr.		-62%	0%	-33%	-54%

Table 10. Results on test data using the character-level system in terms of CER (along with the detailed number of editions).

increase of up to 1.7% on the test data, and a reduction of 50% of the character error rate, which is a very promising result. Those demonstrate the effectiveness of the approach.

5 Discussion

5.1 Systems comparaison

In order to compare results of our two SMT systems, we post-processed the character-level output system to calculate word-level BLEU and WER, and we did the same thing for the word-level output system to calculate character-level BLEU and WER. These scores are detailed in Table 11. As we can see the two SMT systems improves the automatic score of the baseline (OCR-Base), for all conditions (word- and character-level). The best improvements are obtained with the word-level system. Example of errors corrections (translations) of OCR-Base output with the word-level system can be found in Table 12. It can be observed

System	Word level		Char. level	
	BLEU	WER	BLEU	CER
OCR-Base	90.65	4.9	97.39	1.3
SMT Word	96.69	1.9	98.60	0.6
SMT Char.	96.44	2.2	98.59	0.6

Table 11. Comparative results obtained with the two SMT systems at word and character level.

OCR output	Ils viennent , **discnW** - **ls** , vous apporter la paix
SMT Word	Ils viennent , **discnW** - ils , vous apporter la paix
SMT Char	I l s # v i e n n e n t # , # d i s e n t # - # i l s # , # v o u s # a p p o r t e r # l a # p a i x #
REF	Ils viennent , disent - ils , vous apporter la paix
Image	 — Et qu'ils auront fait part à deux.
OCR output	Et **iju**'ils **âiff orit lait pari** à deux .
SMT Word	Et qu' ils **âiff** ont fait part à deux .
REF	Et qu' ils auront fait part à deux .

Table 12. Examples of OCR error correction using SMT system (word-level).

that the system can correct many misrecognized words which are very close to existing words in the vocabulary (*e.g.*"fait" instead of "lait" or "pari" instead of "part") but it has difficulties with some unknowns words (*e.g.*"dicnW" instead of "disent" or "âiff orit" instead of "auront"). This is one of the main shortcoming of the word-level approach. If an error has not been encountered in the training corpus, then the translation model will consider it as an out-of-vocabulary word which cannot be corrected. This kind of error motivates the use of a character-level SMT system, which can overcome this problem by considering a fixed and reduced vocabulary size as well as a longer context.

5.2 Analysis

In order to better understand the impact of the error correction process, an analysis of the correction performed by the systems has been performed. The results are presented in the following. The word-level system modifies 1361 sentences of the test corpus (containing 3004 sentences). For 1329 (97.6%) of them, the number of errors is reduced. Also 12 sentences have their total number of errors unchanged (but error types vary). This may happen when one error is corrected, but another is added. One of the main difficulty of post-processing error correction methods is to be able to modify incorrect hypotheses while not impacting the correct ones. In our case, only 20 sentences (1.5%) are degraded in the process, which is an interesting result.

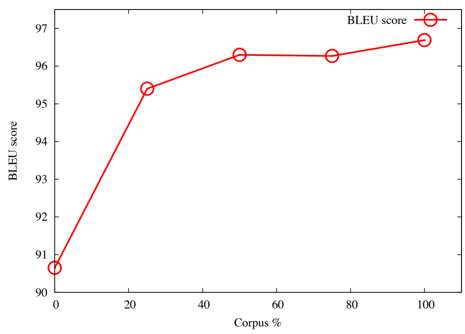

Fig. 2. BLEU score on test using SMT systems trained on different size of corpus.

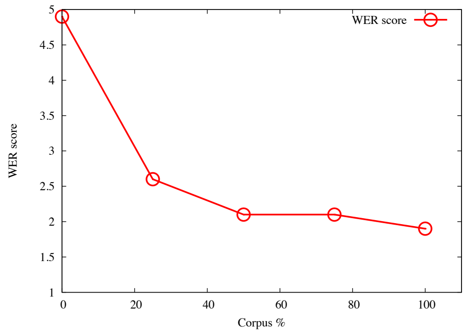

Fig. 3. WER score on test using SMT systems trained on different size of corpus.

In order to analyze the degree of the agreement between the two systems, we scored the character-level system output transformed on word compared to the output of the word-level system. We obtained a BLEU score of 96.55 and WER score of 1.9. This suggest to apply combination of the two systems. This corresponds to correct (translate) the output of the character-level system by the word-level system. As a consequence of this experiment, the BLEU score on word-level evaluation is increased to 97.32 and the WER score is decreased to 1.7 on the same level of the evaluation. This result shows that the two system can complete each other.

6 Conclusion

In this paper, we have shown that using statistical machine translation to correct OCR outputs is feasible and can provide great improvements. The use of this approach allows the system to correct misrecognized words by the OCR. We explored different set-up like character and word-level systems in our experiments. We have shown that such systems are able to (sometimes greatly) improve the BLEU and the WER scores. Our best model outperforms the OCR-Base (baseline) by up to 6 BLEU points and 3 WER points (when computed at word-level), which represents a significant improvement. The results show the superiority of the word-based system compared to character-based system. Nowadays many projects deals with some complex morphological languages like Chinese or Arabic which most of their characters are connected and their shape vary with the position in the word. The morphological complexity of such languages, which have billions surface forms (*e.g.*60 billions for Arabic), complicates others correction methods like dictionary-based [10]. This is mainly because accounting for and listing all the possible words is not an easy task. That is why we believe that our new method can be a good way to resolve this kind of problems. We plan to test it on other different languages and types of data. As future work, we would like to investigate the robustness of our word-level system with respect to domain shift (other domains than historical documents) and for other languages pairs.

7 Acknowledgment

This work has been partially funded by the French Government under the project PACTE.

References

1. 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*, 2012.
2. 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, June 1993.
3. Michel Divay and Anthony J. Vitale. Algorithms for grapheme-phoneme translation for english and french: Applications for database searches and speech synthesis. *Comput. Linguist.*, 23(4):495–523, December 1997.
4. 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, 2008.
5. Tao Hong. *Degraded Text Recognition Using Visual and Linguistic Context*. PhD thesis, University of New York, 1995.
6. Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. *Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, NAACL '07*, pages 372–379, 2007.
7. 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*, ACL '07, pages 177–180, 2007.
8. 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*, NAACL '03, pages 48–54, 2003.
9. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, pages 10:707–710, 1966.
10. 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*, EMNLP '06, pages 408–414, 2006.
11. Preslav Nakov and Jörg Tiedemann. 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, 2012.

12. Kai Niklas. *Unsupervised Post-Correction of OCR Errors*. PhD thesis, Leibniz University Hannover, 2010.
13. Hisao Niwa, Kazuhiro Kayashima, and Yasuham Shimeki. Postprocessing for character recognition using keyword information. *IAPR Workshop on Machine Vision Applications*, 1992.
14. 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*, ACL '03, pages 160–167, 2003.
15. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, 2002.
16. 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*, ACL '12, pages 104–109, 2012.
17. A. Stolcke. SRILM - an extensible language modeling toolkit. In *International Conference on Spoken Language Processing*, pages 257–286, November 2002.
18. Jörg Tiedemann. Character-based pivot translation for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 141–151, 2012.
19. Jörg Tiedemann and Preslav Nakov. Analyzing the use of character-level translation with sparse and noisy datasets. *Proceedings of Recent Advences in Natural Language Processing*, pages 676–684, 2013.
20. Mikael Tillenius. Efficient generation and ranking of spelling error corrections. *NADA REPORT TRITA-NA-E9621*, 1996.
21. David Vilar, Jan-T. Peter, and Hermann Ney. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 33–39, 2007.