Automated Error Detection in Digitized Cultural Heritage Documents

Kata Gábor, Benoît Sagot

Automated Error Detection in Digitized Cultural Heritage Documents

Kata Gábor
INRIA & Université Paris 7
Domaine de Voluceau - BP 105
78153 Le Chesnay Cedex
FRANCE
kata.gabor@inria.fr

Benoît Sagot
INRIA & Université Paris 7
Domaine de Voluceau - BP 105
78153 Le Chesnay Cedex
FRANCE
benoit.sagot@inria.fr

Abstract

The work reported in this paper aims at performance optimization in the digitization of documents pertaining to the cultural heritage domain. A hybrid method is proposed, combining statistical classification algorithms and linguistic knowledge to automatize post-OCR error detection and correction. The current paper deals with the integration of linguistic modules and their impact on error detection.

1 Introduction

Providing wider access to national cultural heritage by massive digitization confronts the actors of the field with a set of new challenges. State of the art optical character recognition (OCR) software currently achieve an error rate of around 1 to 10% depending on the age and the layout of the text. While this quality may be adequate for indexing, documents intended for reading need to meet higher standards. A reduction of the error rate by a factor of 10 to 100 becomes necessary for the diffusion of digitized books and journals through emerging technologies such as e-books. Our paper deals with the automatic post-processing of digitized documents with the aim of reducing the OCR error rate by using contextual information and linguistic processing, by and large absent from current OCR engines. In the current stage of the project, we are focusing on French texts from the archives of the French National Library (Bibliothèque Nationale de France) covering the period from 1646 to 1990.

We adopted a hybrid approach, making use of both statistical classification techniques and linguistically motivated modules to detect OCR errors and generate correction candidates. The technology is based on a symbolic linguistic preprocessing, followed by a statistical module which adpots the noisy channel model (Shannon, 1948). Symbolic methods for error correction allow to target specific phenomena with a high precision, but they typically strongly rely on presumptions about the nature of errors encountered. This drawback can be overcome by using the noisy channel model (Kernighan et al., 1990; Brill and Moore, 2000; Kolak and Resnik, 2002; Mays et al., 1991; Tong and Evans, 1996). However, error models in such systems work best if they are created from manually corrected training data, which are not always available. Other alternatives to OCR error correction include (weighted) FSTs (Beaufort and Mancas-Thillou, 2007), voting systems using the output of different OCR engines (Klein and Kope, 2002), textual alignment combined with dictionary lookup (Lund and Ringger, 2009), or heuristic correction methods (Alex et al., 2012). While correction systems rely less and less on pre-existing external dictionaries, a shift can be observed towards methods that dinamically create lexicons either by exploiting the Web (Cucerzan and Brill, 2004; Strohmaier et al., 2003) or from the corpus (Reynaert, 2004).

As to linguistically enhanced models, POS tagging was succesfully applied to spelling correction (Golding and Schabes, 1996; Schaback, 2007). However, to our knowledge, very little work has been done to exploit linguistic analysis for post-OCR correction (Francom and Hulden, 2013). We propose to apply a shallow processing module to detect certain types of named entities (NEs), and a POS tagger trained specifically to deal with NE-tagged input. Our studies aim to demonstrate that linguistic preprocessing can efficiently contribute to reduce the error rate by 1) detecting false corrections proposed by the

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pages 56–61,
Gothenburg, Sweden, April 26 2014.

statistical correction module, 2) detecting OCR errors which are unlikely to be detected by the statistical correction module. We argue that named entity grammars can be adapted to the correction task at a low cost and they allow to target specific types of errors with a very high precision.

In what follows, we present the global architecture of the post-OCR correction system (2), the named entity recognition module (3), as well as our experiments in named entity-aware POS tagging (4). The predicted impact of the linguistic modules is illustrated in section 5. Finally, we present ongoing work and the conclusion (6).

2 System Architecture

Our OCR error detection and correction system uses a hybrid methodology with a symbolic module for linguistic preprocessing, a POS tagger, followed by statistical decoding and correction modules. The SxPipe toolchain (Sagot and Boullier, 2008) is used for shallow processing tasks (tokenisation, sentence segmentation, named entity recognition). The NE-tagged text is input to POS tagging with MElt-h, a hybrid version of the MElt tagger (Denis and Sagot, 2010; Denis and Sagot, 2012). MELT-h can take both NE tagged texts and raw text as input.
The decoding phase is based on the noisy channel model (Shannon, 1948) adapted to spell checking (Kernighan et al., 1990). In a noisy channel model, given an input string *s*, we want to find the word *w* which maximizes $P(w|s)$. Using Bayes theorem, this can be written as:

$$argmax(w)P(s|w) * P(w) \quad (1)$$

where P(w) is given by the language model obtained from clean corpora. Both sentence-level (Tong and Evans, 1996; Boswell, 2004) and word-level (Mays et al., 1991) language models can be used. $P(s|w)$ is given by the error model, represented as a confusion matrix calculated from our training corpus in which OCR output is aligned with its manually corrected, noiseless equivalent. The post-correction process is summarized in 1.
The integration of a symbolic module for NE recognition and the use of part of speech and named entity tags constitute a novel aspect in our method. Moreover, linguistic preprocessing allows us to challenge tokenisation decisions prior to and during the decoding phase (similarly to Kolak (2005)) ; this constitutes a significant feature as OCR errors often boil down to a fusion or split of tokens.

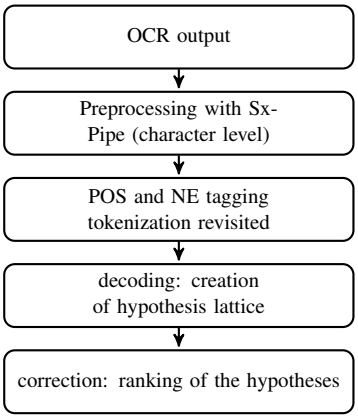

Figure 1: Architecture

The corpus we use comes from the archives of the French National Library and contains 1 500 documents (50 000 000 tokens). This corpus is available both as a "reference corpus", i.e., in a manually corrected, clean version, and as a "contrast corpus", i.e., a noisy OCR output version. These variants are aligned at the sentence level.

3 Named entity tagging

3.1 NE recognition methodology

As a first step in error detection, the OCR output is analysed in search of "irregular" character sequences such as named entities. This process is implemented with SxPipe (Sagot and Boullier, 2008), a freely available[1], robust and modular multilingual processing chain for unrestricted text. SxPipe contains modules for named entity recognition, tokenization, sentence segmentation, non-deterministic multi-word expression detection, spelling correction and lexicon-based patterns detection. The SxPipe chain is fully customizable with respect to input language, domain, text type and the modules to be used. Users are also free to add their own modules to the chain.

In accordance with our purposes, we defined named entities as sequences of characters which cannot be analysed morphologically or syntactically, yet follow productive patterns. Such entities do not adhere to regular tokenization patterns

[1]https://gforge.inria.fr/projects/lingwb/

since they often include punctuation marks, usually considered as separators. As compared to the consensual use of the term (Maynard et al., 2001; Chinchor, 1998; Sang and Meulder, 2003), our definition covers a wider range of entities, e.g., numerals, currency units, dimensions.[2] The correct annotation of these entities has a double relevance for our project:

- NE tagging prior to POS tagging helps to improve the accuracy of the latter.
- NE tagging allows to detect and, eventually, correct OCR errors which occur inside NEs. Conversely, it can also contribute to detect false correction candidates when the sequence of characters forming the NE would otherwise be assigned a low probability by the language model.

The named entity recognition module is implemented in Perl as a series of local grammars. Local grammars constitute a simple and powerful tool to recognize open classes of entities (Friburger and Maurel, 2004; Maynard et al., 2002; Bontcheva et al., 2002); we are concerned with time expressions, addresses, currency units, dimensions, chemical formulae and legal IDs. Named entity grammars are applied to the raw corpus before tokenization and segmentation. Our grammars are robust in the sense that they inherently recognize and correct some types of frequent OCR errors in the input.[3] SxPipe's architecture allows to define an OCR-specific correction mode as an input parameter and hence apply robust recognition and correction to noisy output, while requiring exact matching for clean texts. However, maximizing precision remains our primary target, as a false correction is more costly than the non-correction of an eventual error at this stage. Therefore, our grammars are built around unambiguous markers.

3.2 Evaluation of NE tagging

A manual, application-independent evaluation was carried out, concentrating primarily on precision for the reasons mentioned in 3. For four types of NEs, we collected a sample of 200 sentences expected to contain one or more instances of the given entity category, based on the presence of category-specific markers (lexical units, acronyms etc.)[4]. However, chemical formulae were evaluated directly on sentences extracted from the archives of the European Patent Office; no filtering was needed due to the density of formulae in these documents.
Legal IDs were evaluated on a legal corpus from the Publications Office of the European Union, while the rest of the grammars were evaluated using the BNF corpus.

Entity Type	Precision	Recall
DATE	0.98	0.97
ADDRESS	0.83	0.86
LEGAL	0.88	0.82
CHEMICAL	0.94	-

Table 1: Evaluation of NE grammars

4 POS tagging

4.1 MElt$_{\mathrm{FR}}$ and MElt-h

The following step in the chain is POS tagging using a named entity-aware version of the MElt tagger. MElt (Denis and Sagot, 2010; Denis and Sagot, 2012) is a maximum entropy POS tagger which differs from other systems in that it uses both corpus-based features and a large-coverage lexicon as an external source of information. Its French version, MElt-FR was trained on the Le*fff* lexicon (Sagot, 2010) and on the French TreeBank (FTB) (Abeillé et al., 2003). The training corpus uses a tagset consisting of 29 tags. MElt$_{\mathrm{FR}}$ yields state of the art results for French, namely 97.8% accuracy on the test set.

In order to integrate MElt into our toolchain, the tagger needed to be trained to read NE-tagged texts as output by SxPipe. We thus extended the FTB with 332 manually annotated sentences (15 500 tokens) containing real examples for each type of NE covered by our version of SxPipe. SxPipe's output format was slightly modified to facilitate learning: entities covered by the grammars were replaced by pseudo-words corresponding to their category. The training corpus is the union

[2] Our current experiments do not cover single-word proper names.

[3] E.g., A numerical 0 inside a chemical formula is presumed in most cases to be an erroneous hypothesis for alphabetical O.

[4] Although this sampling is biased towards entities with a certain type of marker, it gives an approximation on the recall, as opposed to simply extracting hits of our grammars and evaluating only their precision.

of the FTB and the small NE corpus annotated with 35 categories (29 POS and 6 named entity categories). We used this corpus to train MElt-h, a hybrid tagger compatible with our OCR post-processing toolchain. MElt-h can tag both raw corpora (using the 29 POS categories learnt from the FTB), and NE-annotated texts (preprocessed with SxPipe or any other tool, as long as the format is consistent with the output of SxPipe).
Training a tagger on a heterogeneous corpus like the one we used is theoretically challengeable. Therefore, careful attention was paid to evaluating it on both NE-annotated data and on the FTB test corpus. The latter result is meant to indicate whether there is a decrease in performance compared to the "original" MElt_{FR} tagger, trained solely on FTB data.

4.2 Evaluation of POS and NE tagging

A set of experiments were performed using different sections of the NE-annotated training data. First, we cut out 100 sentences at random and used them as a test corpus. From the rest of the sentences, we created diverse random partitionings using 50, 100, 150 and all the 232 sentences as training data. We trained MElt-h on each training corpus and evaluated it on the test section of the FTB as well as on the 100 NE-annotated sentences.

#sentences	Prec on FTB	Prec on PACTE-NE
0	97.83	—
50	97.82	95.61
100	97.80	95.71
150	97.78	95.76
200	97.78	95.84
232	97.75	96.20

Table 2: Evaluation of MElt-h on the FTB and on the NE-annotated corpus

The results confirm that adding NE-annotated sentences to the training corpus does not decrease precision on the FTB itself. Furthermore, we note that the results on the NE corpus are slightly inferior to the results on the FTB, but the figures suggest that the learning curve did not reach a limit for NE-annotated data: adding more NE-annotated sentences will probably increase precision.

5 Expected impact on OCR error reduction

While the major impact of named entity tagging and NE-enriched POS tagging is expected to result from their integration into the language model, series of experiments are currently being carried out to estimate the efficiency of the symbolic correction module and the quantity of the remaining OCR errors inside named entities. A sample of 500.000 sentences (15.500.000 tokens) was extracted from the BNF corpus to be used for a comparison and case studies, both in the noisy OCR output version and in the editorial quality version. Both types of texts were tagged for NEs with SxPipe, using the "clean input" mode (without tolerance for errors and correction candidates). Only 65% of the recognized NEs are identical, implying that 35% of the named entities are very likely to contain an OCR error.[5] To investigate further, we applied the grammars one by one in "noisy input" mode. This setting allows to detect certain types of typical OCR errors, with an efficiency ranging from 0 (no tolerance) to 10% i.e., up to this quantity of erroneous input can be detected and correctly tagged with certain named entity grammars. Detailed case studies are currently being carried out to determine the exact precision of the correction module.

6 Conclusion and Future Work

We described an architecture for post-OCR error detection in documents pertaining to the cultural heritage domain. Among other characteristics, the specificity of our model consists in a combination of linguistic analysis and statistical modules, which interact at different stages in the error detection and correction process. The first experiments carried out within the project suggest that linguistically informed modules can efficiently complement statistical methods for post-OCR error detection. Our principal future direction is towards the integration of NE-enriched POS tagging information into the language models, in order to provide a finer grained categorization and account for these phenomena. A series of experiences are planned to be undertaken, using different combinations of token-level information.

[5] In the less frequent case, divergences can also be due to errors in the editorial quality text.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks*. Kluwer, Dordrecht.

Bea Alex, Claire Grover, Ewan Klein, and Richard Tobin. 2012. Digitised historical text: Does it have to be mediOCRe? In *Proceedings of KONVENS 2012 (LThist 2012 workshop)*, pages 401–409, Vienna, Austria.

Richard Beaufort and Céline Mancas-Thillou. 2007. A weighted finite-state framework for correcting errors in natural scene OCR. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 889–893, Washington, DC, USA.

Kalina Bontcheva, Marin Dimitrov, Diana Maynard, Valentin Tablan, and Hamish Cunningham. 2002. Shallow methods for named entity coreference resolution. In *Proceedings of the TALN 2002 Conference*.

Dustin Boswell. 2004. Language models for spelling correction. *CSE*, 256.

Eric Brill and Robert C. Moore. 2000. An Improved Error Model for Noisy Channel Spelling Correction. In *Proceedings of the 38th ACL Conference*, pages 286–293.

Nancy Chinchor. 1998. Muc-7 named entity task definition. In *Seventh Message Understanding Conference (MUC-7)*.

Silviu Cucerzan and Eric Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, pages 293–300, Barcelona, Spain.

Pascal Denis and Benoît Sagot. 2010. Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. In *Traitement Automatique des Langues Naturelles : TALN 2010*, Montréal, Canada.

Pascal Denis and Benoît Sagot. 2012. Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, 46(4):721–736.

Jerid Francom and Mans Hulden. 2013. Diacritic error detection and restoration via part-of-speech tags. In *Proceedings of the 6th Language and Technology Conference*.

Nathalie Friburger and Denis Maurel. 2004. Finite-state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, 313:94–104.

Andrew Golding and Yves Schabes. 1996. Combining trigram-based and feature-based methods for context-sensitive spelling correction. In *ACL*, pages 71–78.

Mark Kernighan, Kenneth Church, and William Gale. 1990. A spelling correction program based on a noisy channel model. In *Proceedings of the 13th conference on Computational linguistics*, pages 205–210.

Samuel Klein and Miri Kope. 2002. A voting system for automatic OCR correction. In *Proceedings of the Workshop On Information Retrieval and OCR: From Converting Content to Grasping Meaning*, pages 1–21, Tampere, Finland.

Okan Kolak and Philip Resnik. 2002. OCR error correction using a noisy channel model. In *Proceedings of the Second International Conference on Human Language Technology Research (HLT'02)*, pages 257–262, San Diego, USA.

Okan Kolak and Philip Resnik. 2005. OCR post-processing for low density languages. In *Proceedings of the HLT-EMNLP Conference*, pages 867–874.

William Lund and Eric Ringger. 2009. Improving optical character recognition through efficient multiple system alignment. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'09)*, pages 231–240, Austin, USA.

Diana Maynard, Valentin Tablan, Cristian Ursu, Hamish Cunningham, and Yorick Wilks. 2001. Named entity recognition from diverse text types. In *In Proceedings of the Recent Advances in Natural Language Processing Conference*, pages 257–274.

Diana Maynard, Valentin Tablan, Hamish Cunningham, Cristian Ursu, Horacio Saggion, Kalina Bontcheva, and Yorick Wilks. 2002. Architectural elements of language engineering robustness. *Journal of Natural Language Engineering - Special Issue on Robust Methods in Analysis of Natural Language Data*, 8:257–274.

Eric Mays, Fred Damerau, and Robert Mercer. 1991. Context based spelling correction. *Information Processing and Management*, 23 (5):517–522.

Martin Reynaert. 2004. Multilingual text induced spelling correction. In *Proceedings of the Workshop on Multilingual Linguistic Ressources (MLR'04)*, pages 117–117.

Benoît Sagot and Pierre Boullier. 2008. SxPipe 2: architecture pour le traitement pré-syntaxique de corpus bruts. *Traitement Automatique des Langues*, 49(2):155–188.

Benoît Sagot. 2010. The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *Proceedings of LREC 2010, La Valette, Malte*.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *In Proceedings of Computational Natural Language Learning*, pages 142–147. ACL Press.

Johannes Schaback. 2007. Multi-level feature extraction for spelling correction. In *IJCAI Workshop on Analytics for Noisy Unstructured Text Data*, pages 79–86.

Claude Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27 (3):379–423.

Christan Strohmaier, Cristoph Ringlstetter, Klaus Schulz, and Stoyan Mihov. 2003. Lexical postcorrection of OCR-results: the web as a dynamic secondary dictionary? In *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03)*, page 11331137, Edinburgh, Royaume-Uni.

Xiang Tong and David Evans. 1996. A statistical approach to automatic OCR error correction in context. In *Proceedings of the Fourth Workshop on Very large Corpora*, pages 88–100.