Post OCR Correction of Swedish Patent Text

The Difference between Reading Tongue 'Lästunga' and Security Tab 'Låstunga'

Linda Andersson[1], Helena Rastas[2], and Andreas Rauber[1]

[1] Vienna University of Technology, Austria
[2] Uppdragshuset AB, Sweden
{andersson,rauber}@ifs.tuwien.ac.at,
helenarastas@uppdragshuset.se

The purpose of this paper is to compare two basic post-processing algorithms for correction of optical character recognition (OCR) errors in Swedish text. One is based on language knowledge and manual correction (lexical filter); the other is based on a generic algorithm using limited language knowledge in order to perform corrections (generic filter). The different methods aim to improve the quality of OCR generated Swedish patent text. Tests are conducted on 7,721 randomly selected patent claims generated by different OCR software tools. The OCR generated and automatically corrected (by the lexical or generic filter) texts are compared with manually corrected ground truth. The preliminary results indicate that the OCR tools are biased to different characters when generating text and the language knowledge of post correction influences the final results.

Keywords: Optical character recognition OCR, error correction algorithm, manual error correction.

1 Introduction

When conducting patent search it is essential that the content in older paper documents are converted into electronic format, since patent search requires high performance in recall [1]. Originally printed material which has been digitized includes errors introduced by the deficiency of the OCR software. The alternative digitization method–manual typing, which has a low error rate for a skilled typist–is too time-consuming and costly when converting a large collection of documents such as patents.

The intention of the OCR process is to extract a full and perfect transcription of the textual content of document images [2]. In the mid-1990s the Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas conducted a series of annual tests of OCR accuracy in order to visualize the capabilities and identify the problems with the state-of-the-art OCR software [3]. The results showed that 20% of the pages contributed about 80% of the errors due to poor page quality. Today, OCR

D. Lamas et al. (Eds.): IRFC 2014, LNCS 8849, pp. 1–9, 2014.

software tools generally have a character accuracy of 99.99% [2]. However, most tests have been performed on English text which according to Nagy is "blessed with one of the simplest scripts in the world" (p. 38) [2]. Furthermore, the OCR software is still not able to provide a high accuracy across heterogeneous document collections. Consequently, when converting legal text, such as patents, post-correction is an essential part of the digitization [4].

In information retrieval (IR) a few OCR errors do not affect the performance when using fuzzy matching, and also IR systems using exact matches can handle a few OCR errors [5]. However, when the collection contains few documents or the documents are short or consist of many low frequent words, the performance will decrease considerably [6].

This paper is organized as follows. Section 2 briefly overviews previous work, genre and language characteristics; Section 3 presents the material and method used in the experiment; Section 4 portrays the results; and Section 5 gives some conclusions and future work.

2 Related Work

Due to the recall demand of patent search it is crucial to have a high accuracy level of OCR read text in order to avoid further vocabulary mismatch. In patent search, even if few documents are responsible for most OCR errors, they can still cause catastrophic effects since finding relevant documents is central before granting a patent –in order to rule out that there exists no prior art [9].

For text mining applications e.g. machine translation (MT), Name entity recognition (NER), the analysis could collapse if the text is not correct.

For the abstract of the patent document EP-1099640-A1anincorrect context language model was selected during the OCR-process. In this case Swedish was mistaken for German and almost all /å/ was interpreted into /ä/.

Föreliggande uppfinning avser en ***uppsamlingsbehällare***för gods av papp, företrädesvis wellpapp, vilken har ***två***par av ***motstäende*** sidoväggar och en botten ***bestäende*** av med sidoväggarna sammanhängande bottenflikar. Behållaren är avsedd att ställas ***pä*** och kombineras med en lastpall, företrädesvis av papp och av ***engängstyp*** och som innefattar ett lastdäck som anligger mot ett antal parallella basbalkar och är förenat med dessa medelst***pä***lastdäckets undersida utformade **längsgäende** och mot basbalkarna vinkelrätt anordnade utskott. ***Lästungor***(14) är anordnade i behållarens (2) pappämne, **ätminstone**vid ett av paren av ***motstäende*** sidoväggar (4, 6), varvid dessa **lästungor** (14) är utformade att ***läsa*****uppsamlingsbehällaren** (2) till lastpallen (30) genom att anligga mot lastdäckets undersida mellan bredvid varandra liggande utskott (36) och genom att inskjutas mellan lastdäck och den närmast liggande basbalken (34) ***sä***att ***lästungan*** därigenom **läser uppsamlingsbehällaren** (2) **pä** lastpallen (30) när denna belastas av godset i ***uppsamlingsbehällaren*** (2). Uppfinningen avser även ett sätt att förankra en ***uppsamlingsbehällare*** för gods vid en lastpall.

Fig. 1. Swedish abstract from patent EP-1099640-A1

The two most important words have been incorrectly processed: 1) ‘uppsamlings-behällare’ a non-word error (see section 2.1), should be ‘uppsamlingsbehållare’ (container for holding goods), 2) ‘lästung,-an,-or’ (reading tongue) a real-word error (see section 2.1), should be ‘låstung,-an,-or’ (security tab). Since the word ‘låstung’ is incorrectly identified in the text, the text gets a second reading entirely different from the original reading. The second reading claims an invention on a reading tongue while the original invention claims an invention on specific type of security tabs.

2.1 Post Processing of OCR Data

The post-process aims to correct the errors that arise due to misinterpretations of graphical similarities during the OCR process [8]. If the image is not clear enough the OCR device either generates a default character (for example ~) or a wrongly identified character or string. Usually the OCR errors are divided into two primary groups: non-word errors – a character string that does not constitute a word in the language or real-word errors – a character string that is a word in the language but does not correspond to the original text.

The previous post-processing research projects can be divided into two main categories: manual and automatic. In manual post-processing different types of interactive corrections of preliminary machine recognition results have been used, mostly relying on volunteers [9]. But the use of volunteers for correction is especially time-consuming [10]. In automatic post-processing, a variety of techniques have been explored, such as language knowledge including lexicon, morphologic, syntactic and semantic information [10]. For statistical language model (SLM), the most used models are the word n-gram models and distance-m n-gram models.

Nylander [10] presented two semi-automatic correction models without the use of a lexicon for Swedish OCR generated text. The first method consists of rules based on n-gram statistics from a training set, and the second method uses a graphotax (i.e. rules for acceptable letter sequences in a language) created from a model of Swedish phonetics.

2.2 Evaluation

Most evaluation methods for OCR software require a ground truth collection. However, overcoming a ground truth is both difficult and expensive. In spite of this, it is essential to perform automatic evaluation of OCR processed documents [11]. There are different ways to establish a ground truth. The most time consuming and expensive is to let several typists type in each single document [3]. Other methods are manual post-correction of OCR generated text and noise model simulation. The latter adds synthetically generated noise to electronic documents [11].

For this experiment we use the ISRI Analytic Tools since the documents in the ISRI ground truth (scientific and technical documents from seven genres produced during a 30-year period) and patent documents have similar structure. The ISRI test research results showed that letters occurring frequently will generally be more correctly recognized than letters occurring less frequently (e.g. text consists of more lowercase than uppercase letters).

2.3 Features of Swedish Patents Text

The Swedish patent documents have three different text sections: abstract, description, claims. Parts of the patent documents are more uniform while others employ a wide variety of type font and type styles. Many also include tables, scientific formulas and other types of graphical material. The sections are not always marked by a headline. Moreover, the headlines are not consistent over the years and the order of the sections alters – the abstract can be at the end of the document as likely as at the beginning. Identifying the ending and beginning of different text sections in order to separate them is not a trivial task. Only identifying page breaks, newlines and changes in size font will not be enough. On newly filed patent document there exist system which conducts identification of document structure according to external information [12].

The Swedish language has some special features that challenge OCR software tools, IR systems and Natural language processing (NLP) applications.

The Swedish alphabet consists of 29 letters. Three additional letters /Å-å/, /Ä-ä/ and /Ö-ö/ complement the 26-letter Modern Roman alphabet.

Swedish morphological units can be subdivided into free morphemes and bound morphemes. One word can contain several free morphemes and bound morphemes, e.g. multi-words which form compounds as orthographic units [13]. A crucial important element for information extraction (NE) application can be concealed inside orthographical compounds [14]. The compound mechanism in Swedish hampers both statistic and lexicon driven methods to correct and detect OCR errors. The inflection hampers pattern matching in IR and affects SLM. For the OCR post-processing this entails that all word forms need to be accounted.

Moreover, the standard recommendation for Swedish text mining application is to use a morphological analysis program for lemmatization and a syntactic parser for ambiguity resolution [15], since Swedish is rich in homographs. Approximately 65 percent of the words in Swedish written text are homographs (e.g. the noun ‘dom’, meaning ‘cathedral’ as well as ‘verdict/ judgment’). The homographs causes many real-word errors generating a second reading entirely different from the original reading, in Figure 1 the word ‘låstunga’ (security tab) was incorrectly identified as ‘läs-tunga’ (reading tongue) a real-word error.

3 Experiment Set Up

Data, The test set for the current experiment consists of machine typed text where emphasis is marked by either underlining a ‘word’ or introducing space between the ‘l e t t e r s’. Out of the 30,217 patent claims in Swedish Claims Collection (SCC)21,746 are found in the manually inspected and corrected part of the Nordic Patents Collection[1] (NPC). The NPC claim texts are used as the ground truth to which the scanned SCC and the scanned Uppdragshuset Claim Collection (UHCC)texts are compared with. Because of limitations in the ISRI Analytic Tools (not being able to deal with documents longer than 60,000 characters) the number of documents is narrowed to

[1]Contributed by Uppdragshuset AB, see `www.nordiskapatent.se`

21,087. A further reduction of number of documents which could be used in the experiment was caused by the issue of correctly identify the being and the end of the claim section in the UHCC, since the original collection contains complete patent documents. Using a simple algorithm, 15,442 claims are collected and every other patent is selected as a sample, leaving 7,721 patent claims to process. The patents range from SE413741 (application filed Dec. 19 1974) to SE438394 (application filed Aug. 15 1979). The average claim in the ground truth selection has 2396 characters (median 1985) and 342 words (284).

3.1 Method

The experiment consists of a comparison between two basic correction processes: Lexical Correction Filter and Generic Correction Filter.

Lexical Filter. The lexical filter was established by extracting 3,000 words which did not retrieve any morphological analysis from the SWETWOL[2] software. The sample was extracted from the SCC collection not used in the experiment. The words were manually assessed and corrected. After optimization 1,142 were found not to conflate with other Swedish words. The space written words are almost fully limited to different forms of the words 'k ä n n e t e c k n a' (characterize) and 'd ä r a v' (that); 114 different pattern of the letter sequence was identified in SCC.

Generic Filter. The generic filter uses a limited amount of lexical resources and observed knowledge outside the material at hand: "it aims to do detection and correction on the fly". The generic filter consists of two algorithms: the first targets space written words, while the second targets OCR errors. The space written word algorithm identifies single letter strings especially targeting the words 'känneteckna' and 'därav'.

The essence of the generic detection and correction algorithm (GDC) is the basic assumption that OCR software tools generates more correct instances of a word than incorrect and that the OCR errors will be inferior to the occurrences of the correct word. The aim of the GDC is the correction of non-word errors, and in order to avoid spurious suggestions constraints are added to the algorithm in terms substitution rules and frequency.

Frequency, words occurring less than 50 times are not used as correction suggestions, only words having a 6% rate of the correction word will be corrected, and words with higher frequency than 100 are not to be corrected. The frequency threshold is based on observations made from 14,775 claims in SCC excluded from the sample set.

Substitution, takes into account position of letters in words, number of differences (only one is allowed), and special letters: if the suggested letter is /ä/ it is allowed to correct /a|å|ö|x|d/; if the suggested letter is /ö/ it is allowed to correct /o/; if the suggested letter is /å/ it is allowed to correct /a/.

[2] `http://www2.lingsoft.fi/cgi-bin/swetwol`

However as mentioned earlier, the inflection, compounding and homographic feature of the Swedish language makes the task difficult. In the current experiment we stemmed all noun suffixes i.e. plural /-or, -ar, -er, -r, -n, Ø,-s/, definitive /-n, -en, -et/ and genitive /-s/ markers [17].

Evaluation . The evaluation consists of manual assessment of the correction suggestions of the generic filter – GDC algorithm. The lexical filter and the generic filter were automatically assessed by the ISRI Analytic Tools. The assessment consisted of computing correctly recognized characters, words and phrases (2-gram to 8-gram). The measurement does not consider extra characters and words inserted by the OCR software tools.

For the manual assessment of the GDC algorithm we asked four native speakers of Swedish to assess the output of the GDC algorithm according to five different criteria:

- if the correction generates a non-Swedish word, select 0;
- if the correction is correct, select 1;
- if both OCR word and correction word are valid Swedish words, select 2;
- if there is another correction suggestion, select 3;
- if the correction still is incorrect, or if it is difficult to assess the word, select 4.

4 Results

The lexical filter, which consists of 1,442 manual correction suggestions, produced 4,555 corrections in SCC while it only corrected 355 words in UHCC. In the SCC material, the generic filter found 535 words which corrected 1,443 instances. For UHCC, the number of words was 73 and corrected 140 instances.

The manual evaluation of the performance of the GDC algorithm produced the following results: 426 words out of the 535 in the SCC material were found to be accurate correction suggestions, while 2 were found to be completely incorrect, e.g. ‘söt’ (‘sweet’) is corrected to ‘*sät’. Almost 107 correction suggestions could not be judged, the words could either be accurate or the error could as likely be corrected by another word.

Out of the 29 documents with the lowest character recognition rate SCC and UHCC share 24. For word recognition 7 of the 11 lowest are mutual. This indicates that there are more problematic documents, for example short patent claims like the patent application SE429652 “Föreningen med formeln” (The compound of formula).

For the automatic assessment we use the ISRI Analytic tool to assess correctly recognized letters, word and phrases.

Table 1 displays the character recognition accuracy per method and data collection.

The results in table 1shows no significant improve for either method or data collection. The generic filter, even, overcorrects giving a negative improvement for UHCC. For SCC the lexical filter marginally improves the accuracy while the generic filter again overcorrects.

Table 1. Correct Recognized special lettersså,ä, ö

Correctly Recognized %	UHCC			SCC		
	OCR	LEXICAL	GENERIC	OCR	LEXICAL	GENERIC
Character	99.46	99.46	99.45	98.97	99.03	98.9
ASCII Uppercase	94.92	94.93	94.44	94.87	94.88	94.64
ASCII Lowercase	99.73	99.73	99.73	99.5	99.52	99.43
Å – å	67.00-99.30	67.00-99.31	67.00-99.30	45.00-95.33	45.00-96.52	45.00-95.65
Ä – ä	82.93-99.49	82.93-99.49	82.93-99.50	87.80-90.29	87.80-92.77	87.80-90.23
Ö – ö	50.57-99.74	50.57-99.74	50.57-99.74	32.95-99.29	32.95-99.31	32.95-99.34

As seen in Table 1, the accuracy of the special character /å/ is marginally improved by the lexical filter in both UHCC and SCC while the generic filter only improves the value in SCC. For /ä/ the generic filter marginally improves the value in UHCC but overcorrect in SCC. For /ö/ the value is only marginally improved in SCC.

The generic filter tend to overcorrects due to the stemming function causing words such as ‘möte’ (meeting) being changed to ‘?mäte’ based on the frequency of ‘*mät derived from the verb ‘mäta’ (measure).

Table 2 shows the accuracy of word and phrases for each method and collection.

Table 2. Correct recognized n-gram sequence

Correctly Recognized %	UHCC			SCC		
	OCR	LEXICAL	GENERIC	OCR	LEXICAL	GENERIC
Unigram	95.46	98.07	95.46	94.14	96.21	95.46
Bigram	91.12	96.3	91.13	88.96	92.8	91.47
Trigram	86.86	94.59	86.88	83.97	89.54	87.65
4-gram	82.67	92.92	82.69	79.25	86.43	84.03
5-gram	78.58	91.29	78.61	74.8	83.46	80.6
6-gram	74.63	89.7	74.67	70.59	80.62	74.28
7-gram	70.92	88.16	70.95	66.62	77.92	74.28
8-gram	67.48	86.67	67.52	62.87	75.36	71.38

The accuracy of words and phrases (see Table 2) are influenced both by the word substitution (lexical filter) or letter substitution (generic filter) but also by the space word matrices (lexical filter) and the word space algorithm (generic filter).

The lexical filter increases the word accuracy from 95.46% to 98.06% in UHCC while the space word algorithm does not have an impact. In SCC both filters increase the correctness for word and phrases, but the performance of the lexical filter is better.

For the generic filter there were several instances where a word with high frequency corrects other word forms of other lemmas. For instance, the word ‘satt’ (with two entirely different meaning i.e. squat and sit) is corrected by the word ‘sätt(a)’ (‘put’) due to the homographic phenomenon. Furthermore, the GDC is not able to correct low frequency compounds with non-word errors, such as correcting ‘*motständskontaktstycket’ to its correct version ‘motståndskontaktstycket’ (motstånds ‘resistance’ # kontakt ‘contact’ # stycke ‘unit’). The lexical filter, on the other hand, is permitted to correct low frequency words since it consists of manually established correction suggestions.

5 Conclusion

In this paper we compare two basic post-process correction filters to improve the quality of OCR generated Swedish patent text. The results show that by applying the recognition of words written with interleaved spaces generates higher accuracy on words and phrases. The result for character accuracy showed that the lexical filter for unseen data (i.e. UHCC) gave no improvement but a slight improvement on unseen data but from the same collection (i.e. SCC), from 98.97 to 99.03 (i.e. SCC). For the generic filter the values decrease under the baseline (only OCR processed) due to the homographic conflation when stemming.

To conclude, even if the manual evaluation of the GDC algorithm showed 426 out 535 was correctly identified by four assessors it needs to be further modified before being used, the next step is to use contextual constraint to handle the over-correction. Furthermore, the GDC algorithm needs to be modified according to the weakness of the OCR software and the frequency threshold should be optimized according to the material. The algorithm needs to handle OCR errors found in low frequency compounds which the lexical filter handles to a certain extent. The results indicate that the stemming used can be harmful towards frequency even if rules are limiting.

In summary, to conduct post-processing on Swedish patent documents is problematic since both domain specific problems (heterogeneous collection – content, vast time period, scanned quality) and language (homographic, inflection and compounding) will affect the final result.

But more importantly, is the lesson learned, if we slightly move from the mainstream genre or change language the language tools and test collections accessible for evaluation become more limited.

References

1. van Dulken, S.: Free patent databases on the Internet: A critical view. WPI 21(4), 253–257 (1999)
2. Nagy, G.: Twenty Years of Document Image Analysis in PAMI. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38–62 (January 2000)
3. Rice, S., Nartker, G.: The ISRI Analytic Tools for OCR Evaluation, UNLV/Information Science Research Institute, TR-96-02 (August 1996)

4. Baird, H.S.: Difficult and Urgent Open Problems in Document Image Analysis for Libraries. In: 6th International Workshop on Document Image Analysis for Libraries, Palo Alto, pp. 25–32 (2004)
5. Vinciarelli, A.: Noisy Text Categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(12), 1882–1895 (2005)
6. Mittendorf, E., Schäuble, P.: Measuring the effects of data corruption on information retrieval. In: 5th Annual Symposium on Document Analysis and Information Retrieval (SDAIR) (1996)
7. Atkinson, K.H.: Toward a more rational patent search paradigm. In: 1st ACM Workshop on Patent Information Retrieval, California, USA, pp. 37–40 (2008)
8. Nylander, S.: Statistics and Graphotactical Rules in Finding. Uppsala University, Dep. of Linguistic (2000)
9. Lin, X.: Quality Assurance in High Volume Document Digitization: A Survey. In: 2nd IEEE International Conference on Document Image Analysis for Libraries, France, pp. 76–82 (2006)
10. Zhuang, L., Zhu, X.: An OCR Post-processing Approach Based on Multi-knowledge. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, pp. 346–352. Springer, Heidelberg (2005)
11. Feng, S., Manmatha, R.: A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. In: 6th Joint Conference on Digital Libraries, pp. 109–118. ACM Press, New York (2006)
12. Boguraev, B.K., Byrd, R.J., Cheng, K.-S.F., Coden, A.R., Tanenblatt, M.A., Wilfried, T.: System and Method for Identifying Document Structure and Assosciated Metainformation and Faclilitating appropriate processing, US 2009/0276378 A1 (November 5, 2009)
13. Teleman, U., Hellberg, S., Andersson, E., Christensen, L.: Svenska Akademiens Grammatik (The grammar of the Swedish Academy), 4 vols. Svenska Akademien, Stockholm (1999)
14. Karlgren, J.: Occurrence of compound terms and their constituent elements in Swedish. In: 15th Nordic Conference on Computational Linguistics, Joensuu, Findland (2005)
15. Hedlund, T.A., Pirkola, A., Järvelin, K.: Aspects of Swedish morphology and Semantics from the perspective of Mono- and Cross-language Information Retrieval. Information Processing and Management 37(1), 147–161 (2001)