Pattern Recognition Letters 35 (2014) 204–213

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Bridging the gap between handwriting recognition and knowledge management

Marcus Liwicki [a,*], Sebastian Ebert [a,b], Andreas Dengel [a,b]

[a] German Research Center for Artificial Intelligence (DFKI) GmbH, Trippstadter Straße 122, 67663 Kaiserslautern, Germany
[b] Knowledge-Based Systems Group, Department of Computer Science, University of Kaiserslautern, P.O. Box 3049, 67653 Kaiserslautern, Germany

ARTICLE INFO

Article history:
Available online 27 December 2012

Keywords:
Handwriting recognition
Information extraction
Ontology
OBIE

ABSTRACT

In this paper we introduce a new layer for the task of handwriting recognition (HWR), i.e., the use of semantic information in form of Resource Description Framework (RDF) knowledge bases. In particular, two novel processing stages are proposed for the first time in literature. The first stage is the inclusion of RDF knowledge bases into the HWR process, where we make use of a person's mental model. This process can be extended to use other ontological resource. The second stage is the transition from pure handwriting recognition to understanding the handwritten notes, i.e., the system extracts knowledge employing RDF knowledge-bases. This is also called ontology-based information extraction (OBIE). The task of our recognizer therefore is not only to recognize the ASCII transcription of the handwritten document, but also to identify the semantic concepts which appear in the text. For both novel approaches we performed a set of experiments on various data. First, the recognition rate of the HWR system is increased on several documents. Second, the performance of information extraction is also remarkable. By using the *k*-best word recognition alternatives in form of a lattice as an input for the OBIE system, the performance reaches a level which is very close to OBIE applied on pure ASCII text.

1. Introduction

Handwriting recognition (HWR) has been the topic of research for many decades. While the first recognizers have been developed for isolated characters or digits, later recognizers focused on complete words or even sentences (Bunke, 2003; Plamondon and Srihari, 2000; Vinciarelli, 2002). Nowadays there exist solutions which have a quite good recognition performance (e.g., recognizers from Microsoft© and Vision Objects©).

However, the task of handwriting recognition cannot be assumed to be solved already. There is still room for improvement for the recognition performance, as well as handling different scripts and special environments. Currently, much research effort goes into the direction of improving recognizers in these use cases (Chaudhuri et al., 2010).

In this paper we go one step further. Instead of just recognizing the handwritten text, we try to understand the meaning of the written content. For many applications not only the ASCII transcription, but also the important content and concepts are of interest. This can be used to categorize the document or even to relate it to other documents and known concepts in the knowledge space of a person or a company.

Considering the process of note-taking, for example, the person would write down newly acquired knowledge about instances which might appear already in his or her personal knowledge space. Our proposed system can extract the information and identify the new knowledge based on the written content. Finally, the user just needs to check the correctly identified information. This would decrease the work-load of the person significantly, because usually this information has to be typed into the computer and formalized manually.

Recent advance in knowledge management allows to extract information from unstructured text which is available in ASCII format (Adrian et al., 2009). The so-called *ontology-based information extraction* (OBIE) (Wimalasuriya and Dou, 2010) relies on general knowledge in form of an ontology. A user-specific knowledge base, for example, can be formalized in an RDF-graph[1] and made available in a Semantic Desktop (Decker and Frank, 2004; Dengel, 2007). OBIE uses this formalized knowledge and identifies the concepts which appear in the handwritten text. Based on this information, new knowledge can be generated, which just needs to be shortly confirmed by the user (instead of typing the new information explicitly by the keyboard). OBIE methods first segment the text into tokens, then identify their values and their corresponding instances of the ontology, and finally try to generate new facts based

* Corresponding author. Tel.: +49 0 631 205 75 1200; fax: +49 0 631 205 75 1020.
E-mail address: marcus.liwicki@dfki.de (M. Liwicki).

[1] The Resource Description Framework is described in <http://www.w3.org/RDF/>.

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.12.012

on the text. To the authors' knowledge, in this paper OBIE is proposed for the first time in handwriting literature.

Our proposed system performs a seamless integration of the handwriting recognition into the OBIE process. Instead of just applying OBIE on the recognized text, we designed an integrated process which also takes the top-*k* alternatives for each word into account. In our experiments we measure the number of correctly extracted instances. We found that considering more than just the top-candidate improved the performance.

Note that this article is an extended version of Ebert et al. (2010). However, while Ebert et al. (2010) focused on the overall system description and experiments, this article gives more background information of ontologies and handwriting recognition. Furthermore, the methods for improving handwriting recognition are described with more detail and experiments for this task are included. Finally, a novel set of experiments is performed on short handwritten notes in order to compare the behavior on handwritten notes to the behavior on handwritten texts.

The paper is organized as follows. First, Section 2 gives an overview of the general structure of HWR systems and introduces two examples that were used in our experiments. Furthermore, related work in the field of information extraction is presented. Second, Section 3 deals with the representation of knowledge and explains basic concepts that are used throughout the paper. Next, Section 4 shows how semantic information can be incorporated in the HWR process to increase the recognition performance and gives experimental results on this approach. Subsequently, Section 5 describes our approach on how to extract knowledge out of handwritten text. Experimental results are also reported in Section 5. Finally, Section 6 concludes the paper and gives directions for future work.

2. Background

2.1. State-of-the-art HWR systems

This section gives an overview about the handwriting recognition system in general and the main contribution of Section 4. The main steps performed in handwriting recognition are illustrated in Fig. 1, they consist of preprocessing, normalization, feature extraction, classification, and finally a postprocessing step.

Preprocessing is the first step in the handwriting recognition system where the noise associated to the sample input is eliminated. This step often comprises line extraction, and sometimes word separation and character segmentation, depending on the recognition task. However, character segmentation is a very difficult problem. On the one hand side, it is not possible to segment a word into characters before recognizing this word and on the other hand side the word cannot be recognized correctly before being segmented into characters. This situation is known as Sayre's paradox (Sayre, 1973).

Normalization decreases the effect of various writing styles by normalizing the input handwritten data. It can also be considered among the previous step. In normalization, the characters' skew, slant, height and width are adjusted.

Feature extraction acquires the set of feature vectors from the input sample. This particular step is needed because the classifier usually needs numerical values as an input instead of using the raw point-sequence data.

Classification is the process where the feature vectors are fed to classifiers like Hidden Markov Models (HMMs) and Neural Networks (NNs) to obtain recognition candidates. Often, multiple alternatives are provided by the recognizer together with a recognition probability.

Postprocessing comprises several steps which can be performed on the recognizer's output. Very often word lexicons or even grammars are used to improve the recognition result.

We use the *Microsoft Handwriting Recognizer*©[2] for parts of our experiments. This recognizer extracts some online and offline features from oversegmented characters and applies TDNN classifier for the recognition. Dictionary information is integrated by using a trie-based approach. For more information about the recognizer, refer to the work of Pittman (2007).

As an alternative, the MyScript recognizer from Vision Objects© was used for the recognition.[3] The overall recognition system is built on the principles presented by Knerr et al. (1997). Furthermore, a state-of-the-art statistical language model as described by Perraud et al. (2006) is used.

The contribution of Section 4 is to enhance the postprocessing by the integration of semantic information. The semantic information is extracted from a representation of the user's mental model. More specific information about the mental models and their representation is given in the next section.

2.2. Related work in information extraction on handwritten documents

The contribution of Section 5 is to extract information from the handwritten notes. Several other research areas are related to this task. Word spotting, for example, is the task of finding a given word in a handwritten text (Manmatha et al., 1996). Usually, the word is presented as a query of the user who wants to find those places where the specific word appears. At first glance word-spotting seems to be similar to ontology-based information extraction, since specific words are to be retrieved. However, in word spotting there is only a single query while ontology-based information extraction tries to find semantic instances given in an ontology, which might be very complex. Furthermore, we do not just apply a search algorithm, instead we also take relations between the concepts in the RDF knowledge base into account.

Another related task is the retrieval of documents out of a given document corpus. Document retrieval became more and more popular in the last years. Here the task is to find (retrieve) or classify a given set of documents (Pena Saldarriaga et al., 2010). Even if

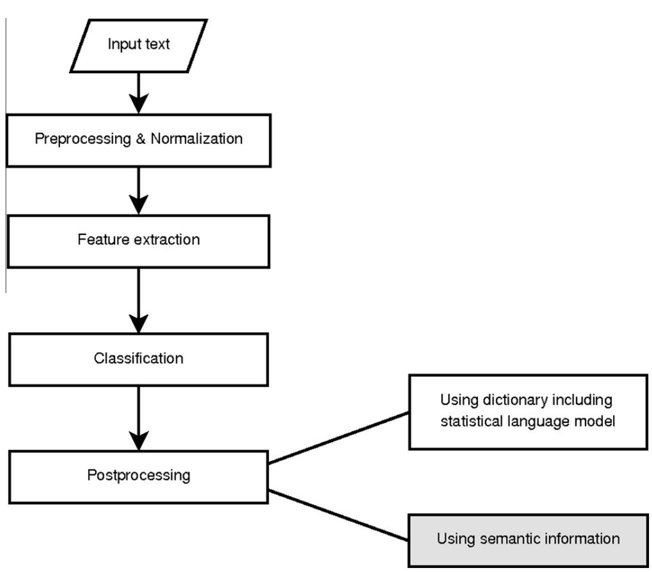

Fig. 1. General handwriting recognition and our main contribution: we include semantic information into the recognition process.

[2] The Microsoft Windows XP Tablet PC Edition SDK© is available for download at <http://www.microsoft.com/windowsxp/tabletpc/default.mspx>.

[3] The MyScript Builder SDK© is available for purchase at <http://www.visionobjects.com/>.