An On-Line Japanese Handwriting Recognition System integrated into an E-Learning Environment for Kanji

Diplomarbeit

zur Erlangung des Grades
eines Diplom-Linguisten
der
Fachrichtung 4.7 Allgemeine Linguistik
der Universität des Saarlandes.

Anfertigt von Steven B. Poggel
sbp@coli.uni-saarland.de

unter Leitung von
Prof. Dr. Dr. h.c. mult. Wolfgang Wahlster
und
Dr. Tilman Becker

Saarbrücken, den 31.03.2010

Contents

Summary

In this work, a Japanese handwriting recognition system is being developed. The system is integrated into an e-learning environment in order to provide a Kanji teaching application with automated error correction. Conceptually, the application is an e-learning environment for Japanese characters, intended for the western learner of the Japanese language. Most e-learning systems of Japanese Kanji provide only a multiple choice method for the learner to reproduce characters. The present prototype offers the ability to enter characters with a stylus on a touch screen system.

The study seeks to determine to what extent it is possible to use modern NLP methods for language learning. While other studies mainly focus on grammatical correction, this application is targeted on the Kanji characters. It will be examined if a handwriting recognition engine can generate informed feedback, suitable for a learner. Additionally, the study examines if that feedback helps obtaining the ability to actively reproduce the Kanji characters.

The prototype developed in this work combines e-learning methods with natural language processing applied to the Japanese script. In order to prepare the task of creating an interdisciplinary software that spans across the aforementioned fields of study, the work reviews the structure of the Japanese script, the current state of the art in handwriting recognition methods and e-learning techniques. The recognition engine implements a structural approach to Kanji character identification. The recognition performs partial analysis of substructures and binds the recognised elements together to form a character. Because of the structural approach it becomes possible to create an informed error recognition that considers linguistic units of the Kanji characters.

The study resulted in the knowledge that it is indeed possible to use a handwriting recognition engine for the generation of informed feedback based on the errors made.

Zusammenfassung

In dieser Arbeit wird eine Handschriftenerkennung für japanische Kanji entwickelt. Der Handschriftenerkenner ist in eine E-Learning-Umgebung integriert und liefert eine automatisch generierte Fehlerkorrektur für Lernende.

Das System ist in konzeptioneller Hinsicht eine E-Learning-Anwendung für das Erlernen der japanischen Schrift. Letztere weist aufgrund ihrer morphemischen Struktur einen hohen Komplexitätsgrad auf und benötigt daher besonderen Lernaufwand. Die meisten E-Learning-Systeme für asiatische Schriftzeichen bieten Zeichenabfrage als Multiple-Choice an, da die Eingabe der Zeichen für einen Lernenden sonst ein technisches Problem darstellen würde. Der in dieser Arbeit erstellte Prototyp bietet die Möglichkeit zur handschriftlichen Eingabe von Zeichen auf einer dafür geeigneten Bildschirmoberfläche. Das ist ein Alleinstellungsmerkmal unter den E-Learning-Anwendungen für die japanische Sprache.

Die Studie untersucht, inwieweit es im Bereich des Schrifterwerbs möglich ist, NLP und Lernmethoden zusammenzubringen. Dabei wird nicht mit Parsing-Methoden die grammatische Struktur der Sprache untersucht, sondern vielmehr die interne Struktur der Kanji zugrunde gelegt und durch einen Handschriftenerkenner erfasst. Dabei sollen Schreibfehler strukturell erkannt werden. Intelligentes Feedback soll dem Lernenden dabei helfen, die Fähigkeit der aktiven Reproduktion der Kanji zu erwerben.

Da der Prototyp eine disziplinübergreifende Software ist, die in den Bereichen Handschriftenerkennung und E-Learning angesiedelt ist, wird in der vorliegenden Arbeit der Forschungsstand der beiden untersucht. Weiterhin wird die Struktur der japanischen Schrift linguistisch analysiert und dargestellt. Die Kombination der drei Disziplinen in einer Studie führt dazu, dass die Substrukturen der Kanji überhaupt programmatisch analysiert werden können, wodurch die Fehlererkennung ermöglicht wird.

Das Ergebnis der Studie zeigt, dass es möglich ist, eine Handschriftenerkennung nutzbar zu machen, um einem Lernenden intelligentes Feedback zu geben, basierend auf den Fehlern, die bei der Eingabe gemacht wurden.

Chapter 1

Introduction

1.1 Motivation

In the history of Computational Linguistics there have been a several attempts to integrate natural language processing techniques with existing technologies. That task is complex in many aspects, depending on what is aimed at exactly.

In this study, I will attempt to create a handwriting recognition for Japanese Kanji. That seems interesting because Kanji the characters form a morphemic writing system with a large number of characters. Thus, handwriting recognition (HWR) of Kanji follows different patterns than in alphabetical writing systems like Latin script. The overall methodology of HWR system is similar to systems for Latin characters, but the details of analysis vary strongly.

Studying Japanese language is a complex task, because a new learner has to get used to a new vocabulary that - coming from a European language - has very little in common with the vocabulary of his mother tongue, unlike in European languages where quite often there are several intersections. The learner also needs to learn a new grammar system. Broadly speaking, most of the central European languages follow a subject-verb-object (SVO) structure. Japanese follows a subject-object-verb (SOV) structure. These create additional difficulty, comparable with German reversed subclause structures that are a source of error for many learners of German. Yet, the most notable difference for a language learner with a central European mother tongue is of course the writing system. The Japanese writing system uses three different scripts. The so-called *Kana* scripts *Hiragana* and *Katakana* are syllabic, each character represents a syllable. Each syllable consists of either a vowel, a consonant and a vowel, or a consonant cluster and a vowel. The syllables are called *open*. Hiragana and Katakana represent roughly the same inventory of syllables and both have around 40-50 characters that can be modified with diacritics in order to yield additional syllable representations. Therefore, these scripts are a hurdle for a learner, but relatively unproblematic, due to their limited number of characters. Besides, the two sets of Kana characters look quite distinct, so the problem of confusing one character with another is limited to a relatively short learning period of those two scripts.

The Kanji, on the contrary, form the largest part of a writing system that has around 3,000 characters, which are built up of around 200 subunits called *Radicals*. One part of the complexity lies in the number of characters. The other part lies in the general method of representing an idea or concept with a character instead of attempting to represent the sounds with graphemes in connection with language specific pronunciation rules. Another difficulty lies in cognitively connecting the characters with their pronunciations. Most characters have multiple pronunciations and for a language learner, studying Japanese vocabulary takes at least twice as much effort compared to languages using a Latin or at least some kind of alphabetic writing system. The two tasks of learning the Kanji and studying the vocabulary together can epitomise a very high learning curve. A connected subordinated problem lies in the fact that quite often subjectively 'simple' vocabulary comes with complex Kanji. Some e-learning applications have taken on that issue by creating a learning environment in which a learner can connect learning vocabulary with studying the Kanji characters.

1.1.1 Integrating NLP and E-Learning

In this project, we would like to approach the issue of studying Kanji in an e-learning application. The novelty about it is a handwriting recognition that gives the learner the ability to actually practise writing the Kanji, instead of the rather limited multiple choice recognition that most other applications use.

List of Figures

Listings

List of Tables

References

Document created on Tuesday 6th April, 2010 at 09:28