Chapter 1

Notes

This project uses state of the art Chinese/Japanese handwriting recognition methods in order to provide an Kanji teaching application with an error correction.

Conceptually, the application is an e-learning environment for Japanese characters, intended for the foreign learner of the Japanese language. In order to provide more than a multiple choice method, like most other systems, the application contains a handwriting recognition engine that can be used preferably with a hand-held device like a PDA, but generally any stylus input device.

The handwriting recognition method used is similar to the one proposed by Chen, J.-W. and S.-Y. Lee (1996) in their article "A Hierarchical Representation for the Reference Database of On-Line Chinese Character Recognition" in "Advances in Structural and Syntactical Pattern Recognition", Volume 1121 of "Lecture Notes in Computer Science", pp. 351--360. Berlin/Heidelberg, Germany: Springer.

(Nakagawa, Tokuno, Zhu, Onuma, Oda, and Kitadai 2008) report their recent results of online Japanese handwriting recognition and its applications. Their article gives important insights into character modeling, which are employed in this application.

Japanese HWR

Steven Buraje Poggel
steven.poggel@gmail.com

November 4, 2009

DFKI
IUI
Prof. Wolfgang Wahlster
Department of Computational Linguistics
Saarland University

Contents

Abstract

Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet. Lorem dolor sit amet. Lorem ipsum dolor sit amet.

Chapter 2

Introduction

2.1 Motivation

In the history of Computational Linguistics there have been a several attempts to integrate natural language processing techniques with existing technologies. This work is one just like that. Concretely, we will try to create a handwriting recognition for Japanese Kanji. That seems interesting, because Kanji is an iconographic writing system, thus handwriting recognition (HWR) can follow different patterns than in alphabetical writing systems like latin.

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 therefore creating additional difficulty, comparable with German subclause structures that are a source of error for 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 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. Hiragana and Katakana represent roughly the same set of syllables and both have around 40-50 characters that can be modified with diacritics and thus yield additional syllable representations. Therefore, these scripts are a hurdle, but relatively unproblematic, due to their limitation in number of characters. Besides, they look quite distinct, so there is the problem of confusing one character with another, but this is limited to a relatively short period of learning those characters.

Kanji, however, is an iconographic writing system that has around 2000 characters, which are built up of around 200 subunits called 'radicals'. So one part of the complexity lies in the number of characters. The other part of the complexity lies in the general concept of representing an idea or concept with a character instead of representing the phonemes of the spoken language with graphemes in connection with some language specific pronunciation rules. Another difficulty lies in connecting the characters with their pronunciations. Most characters have multiple pronunciations and for a language learner, studying Japanese vocabulary is a double or triple task compared to languages using a Latin or at least an alphabetic writing system. Therefore, the two tasks of learning the Kanji and studying the vocabulary together can epitomise a very high learning curve. A subordinated issue connected to that is 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.

2.1.1 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.

2.1.2 Another subsection with a yet unknown title

2.2 A CJK environment

Rather than selecting a CJK font as the main document typeface, you might want to define a CJK environment for text fragments used in the midst of a document using a normal Roman font. This allows me to say `\begin{CJK}東光\end{CJK}` to generate 東光, without putting the whole paragraph into the Far Eastern font. Or I could define a command that takes the CJK text as an argument, so that `\cjk{北京}` produces 北京. It's that easy!

2.3 Running text

コンピューターは、本質的には数字しか扱うことができません。コンピューターは、文字や記号などのそれぞれに番号を割り振ることによって扱えるようにします。ユニコードが出来るまでは、これらの番号を割り振る仕組みが何百種類も存在しました。どの一つをとっても、十分な文字を含んではいませんでした。例えば、欧州連合一つを見ても、そのすべての言語をカバーするためには、いくつかの異なる符号化の仕組みが必要でした。英語のような一つの言語に限っても、一つだけの符号化の仕組みでは、一般的に使われるすべての文字、句読点、技術的な記号などを扱うには不十分でした。

これらの符号化の仕組みは、相互に矛盾するものでもありました。二つの異なる符号化の仕組みが、二つの異なる文字に同一の番号を付けることもできるし、同じ文字に異なる番号を付けることもできるのです。どのようなコンピューターも（特にサーバーは）多くの異なった符号化の仕組みをサポートする必要があります。たとえデータが異なる符号化の仕組みやプラットフォームを通過しても、いつどこでデータが乱れるか分からない危険を冒すことのなるのです。

Chapter 3

Japanese Script

3.1 A Short History of the Japanese Writing System

3.2 The Japanese Writing System Today

kurze erwaehnung der morphologie. hiragana an verben zur konjugation. zusammenhang verben / nomen in kanji.

uppercase / lowercase nicht vorhanden. etc.

3.3 Writing Japanese - typical errors

Chapter 4

On-Line Handwriting Recognition

4.1 General Methods

Handwriting recognition has been around for many years. The first research papers concerned with pattern recognition on computers were published in the late 1950ies, Handwriting recognition as an individual subject in the early 1960ies. (Goldberg 1915) describes a machine that can recognise alphanumeric characters in a US Patent as early as 1915, however that was before the times of modern computers, therefore the methods he employs are different from the algorithms used after the advent of computers, more concretely, computers with screens.

(Tappert et al. 1990) describe in their review the development of handwriting recognition, which was a popular research topic in the early 1970ies and then again in the 1980ies, due to the increased availability of pen-input devices. Generally speaking, handwriting recognition (HWR) involves automatic conversion of handwritten text into a machine readable character encoding like ASCII or UTF-8. Typical HWR-environments include a pen or stylus that is us ed for the handwriting, a touch-sensitive surface, which the user writes on and a an application that interprets the strokes of the stylus on the surface and converts them into digital text. Usually, the writing surface captures the x-y coordinates of the styus movement.

4.2 Hardware requirements

Several different hardware commercial products are available in order to capture the x-y coordinates of a stylus or pen. Graphics tablet like the products of the Wacom Co., Ltd.[1] are popular input devices for hand motions and hand gestures. The use of pen-like input devices has also been recommended, since 42% of mouse users report feelings of weakness, stiffness and general discomfort in the wrist and hand when using the mouse for long periods (Woods et al. 2002). Moreover there are PDAs and Tablet PCs, where the writing suface serves as an output device, i.e. an display at the same time. New generation mobile phones also contain touch-displays, but for those it is more common to be operated without a stylus. Those devices interpret user gestures, however the input is given directly with the users fingers. Another rather new development are real-ink digital pens. With those, a user can write on paper with real ink, and the pen stores the movements of the pen-tip on the paper. The movements are transferred to a computer later. It can be expected that with technologies like Bluetooth it may be possible to transfer those data in real-time, not delayed.

4.3 On-Line vs. Off-Line recogniton

On-line HWR means that the input is converted in *real-time*, *dynamically*, while the user is writing. This recognition can lag behind the user's writing speed. (Tappert et al. 1990) report average writing rates of 1.5-2.5 characters/s for English alphanumerics or 0.2-2.5 characters/s for Chinese characters. In online systems, the data usually comes in as a sequence of coordinate points.

Off-line HWR is the application of a HWR algorithm after the writing. It can be performed at any time after the writing has been completed. That includes recognition of data transferred from the real-ink pens (see 4.2) to a computing device after the writing has been completed. The stadard case of off-line HWR, however, is a subset of optical character recognition (OCR). An scanner tranfers the physical image on paper into a bitmap, the character recognition is performed on the bitmap. An OCR system can recognise several hundred characters per second.

On-line devices have the dynamic information of the writing, since each point coordinate is captured at a specific point of time. Also, the system know the input stroke sequence, their direction and speed of writing. All these information can be an advantage for an on-line system, however, off-line systems have used algorithms of line-thinning, such that the data consits of point coordinates, similar to the input of online systems (Tappert et al. 1990).

[1] www.wacom.com

4.4 Typical On-Line HWR application

A typical HWR application has several parts that follow up on each other in a procedural fashion.

- **Data capturing**: The data is captured through an input device like a writing surface and a stylus.
- **Preprocessing**: The data is segmented, noise reduction like smoothing and filtering are applied.
- **Character Recognition**: Feature analysis, stroke matching, time, direction and curve matching.

4.5 HWR of Hanzi and Kanji

- Warum: Um einen Ueberblick ueber HWR-Techniken fuer Japanische Schriftzeichen und verschiedene Herangehensweisen zu verschaffen.
- Nutzen: Leser kann sich ein Bild darueber verschaffen, in welchem Kontext sich die Applikation bewegt.
- Was: research different approaches, see what the focus on, what their specialty is and report about them
- Wie: Wiss. Report. / Zusammenfassung. Vergleich.

the only subpara and what does this look like?

4.5.1 The current State-of-the-Art in Japanese and Chinese Character Recognition

From the 1990s onwards, On-Line Japanese and Chinese Character Recognition (OJCCR) systems have been aiming at loosening the restrictions imposed on the writer when using an OJCCR system. Their focus shifted from recognition of block style script ('regular' script) to fluent style script, which is also called 'cursive' style. Accuracies of up to about 95% are achieved in the different systems. (Liu, Jaeger, and Nakagawa 2004) have said: bla. says the opposite. (Chen and Lee 1996) oder auch (Nakagawa, Tokuno, Zhu, Onuma, Oda, and Kitadai 2008) und (Nakai, Shimodaira, and Sagayama 2003) zu guter letzt: (Santosh and Nattee 2009)

4.5.2 Overview of a typical OJCCR system

Chapter 5

e-learning

5.1 General E-Learning methods

5.2 E-Learning of languages

in section

5.3 E-Learning of Japanese

5.3.1 Conceptual issues

5.3.2 Japanese e-learning software

put all your bashing and criticism here

Chapter 6

Conceptual design of Kanji-Coach

6.1 General requirements

6.2 Tackling the difficulties of the Japanese script

6.3 Integration of HWR into learning process

6.4 Use cases

Chapter 7

Technical Design of the Application

7.1 System Architecture

7.2 Framework and Devices

7.2.1 Pen Input Device

7.2.2 Desktop Computer

7.3 GUI

7.4 Communication

Chapter 8

Handwriting Recognition Engine

The sections of this chapter are more the result of a brainstorming than a proper thought-through chapter design.

8.1 Capturing Data

this should deal with how the data are captured during the process mouse coordinates and stuff

8.2 Data Format

how is the data structured? radicals, strokes, characters, xml-format

8.3 Database

where did I get it from? how many chars are in there? how are they accessible? what format?

8.4 Recognition Architecture

8.5 Stroke recognition process

8.5.1 From point list to vectors

8.5.2 Handling curves

8.5.3 Handling all that other stuff that requires some math

8.6 Radical recognition process

8.7 Character recognition process

8.8 Error recognition

8.8.1 How to deal with typical errors when writing Japanese

Error recognition

focus on technical aspects

Error handling

focus on technical aspects

8.9 HWR applied to e-learning of Japanese Kanji

8.9.1 Integration of HWR into e-learning app

educational aspects / the e-learning view

8.9.2 Error handling

educational aspects / the e-learning view

Chapter 9

Implementation and Evaluation

9.1 Implementation Details

Pointer auf CD und auf Appendix mit Beispielinteraktionen (diese mit Foto). Screenshots. Zahlen zur Erkennung - z.B. wie lange dauert es, ein zeichen zu erkennen?

9.2 Evaluation of the HWR

9.2.1 Evaluation Metrics

evaluation method: counting precision and recall section about precision and recall - the odd numbers. how can that be done honest and useful? how can I get meaningful numbers at all?

9.3 Evaluation of E-Learning Application with Integrated HWR

qualitative auswertung, keine zahlen, sondern fragebogen. fuehlt der lerner sich unterstuetzt? glaubt er, dass es schneller geht als ohne HWR? besser als auf papier?

9.4 Evaluation of the Error Hints

use cases, inwieweit helfen die fehlerhinweise? geht das lernen dann wirklich schneller? wie laesst sich der mehrwert bewerten? system kann sagen:

wo liegt die verwechslung? warum war das falsch?

Chapter 10

Conclusions

Chapter 11

Outlook

Was kann man hiermit fuers chinesische / koreanische machen? HWR als Service fuer andere Applicationen.

References

Chen, J.-W. and S.-Y. Lee (1996). A Hierarchical Representation for the Reference Database of On-Line Chinese Character Recognition. In *Advances in Structural and Syntactical Pattern Recognition*, Volume 1121 of *Lecture Notes in Computer Science*, pp. 351--360. Berlin/Heidelberg, Germany: Springer.

Goldberg, H. E. (1915, December). Controller. *United States Patent 1,116,663*.

Liu, C.-L., S. Jaeger, and M. Nakagawa (2004). Online Recognition of Chinese Characters: The State-of-the-Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*, 198--213.

Nakagawa, M., J. Tokuno, B. Zhu, M. Onuma, H. Oda, and A. Kitadai (2008). Recent Results of Online Japanese Handwriting Recognition and Its Applications. In D. Doermann and S. Jaeger (Eds.), *Arabic and Chinese Handwriting Recognition*, Volume 4768 of *Lecture Notes in Computer Science*, pp. 170--195. Berlin/Heidelberg, Germany: Springer.

Nakai, M., H. Shimodaira, and S. Sagayama (2003). Generation of Hierarchical Dictionary for Stroke-Order Free Kanji Handwriting Recognition Based on Substroke HMM. In *Proc. Seventh Int'l Conf. Document Analysis and Recognition*, pp. 514--518.

Santosh, K. and C. Nattee (2009). A Comprehensive Survey on On-Line Handwriting Recognition Technology and its Real Application to the Nepalese Natural Handwriting. *Kathmandu University Journal of Sience, Engineering and Technology 6*(I), 30--54.

Tappert, C. C., C. Y. Suen, and T. Wakahara (1990). The State of the Art in Online Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*(8), 787--808.

Woods, V., S. Hastings, P. Buckle, and R. Haslam (2002). *Ergonomics of using a mouse or other non-keyboard input device*, Chapter 3, pp.

23. Number 045 in HSE research report. London: Health and Safety Executive.