PROCEEDINGS OF THE INTERNATIONAL CONFERENCE
ON FIFTH GENERATION COMPUTER SYSTEMS 1992,
edited by ICOT.

A Hybrid Reasoning System For Explaining Mistakes In Chinese Writing

Jacqueline Castaing
Univ. Paris-Nord, Lipn / . Csp., Avenue J-B Clément
93430 Villetaneuse France
jc@lipn.univ-paris13.fr

Abstract

We present in this paper a hybrid reasoning system for Explaining Mistakes In Chinese Writing, called EMICW. The aim of EMICW is to provide students of the chinese language with a means to memorize characters. The students write down from EMICW 's dictation. In case of graphic errors, EMICW will explain the reasons of this error by using either the etymology of characters or some efficient mnemonic techniques.

EMICW has multiple representations associated to multiple reasoning methods. The coherence of the reasoning is ensured by means of a common logic formalism, the FLL-theories, derived from Girard's linear logic.

1 Introduction

The main aim of the system EMICW is to provide students of the chinese language with a means to memorize chinese characters without losing heart. The first obstacle for people accustomed to an alphabet is indeed the great number of characters to sink in. We propose to them to write down from EMICW's dictation. In the case where students are mistaken about a character, the system will explain the reasons of this graphic error either by using the origin of the character [Henshall 1988], [Ryjick 1981], [Wieger 1978], or by invoking an efficient mnemonic technique.

EMICW is a hybrid knowledge representation and reasoning system [Brachman, et al. 1985], [Kazmareck et al. 1986], [Nebel 1988]. It has multiple representations - a semantic network associated to inference rules expressed in the formalism of Gentzen's calculus [Gentzen 1969] - associated to multiple reasoning methods. The set of inference rules defines the main cases of mistakes that the author of this article and school fellows could make during their own initiation into the chinese writing. The learning methods used are given in [Bellassen 1989], [De Francis 1966], [Lyssenko and Weulersse 1987] [Shanghai Press 1982].

To ensure a coherent reasoning, EMICW has a common logic formalism, the FLL-theories [Castaing 1991], borrowed from Girard's linear logic [Girard 1987, 1989]. The system essentially performs monotonic abduction [Bylander 1991]. So, let a be the correct chinese character the student should write down from EMICW 's dictation. Let b be the actual answer given by the student. If the student is mistaken, it means that the character a is different from b, the binary predicate Error (a, b) is then set to the value true. An explanation of a graphic error consists in finding a set of first-order formulas Sigma such that a proof of the linear sequent Sigma ⊢ Error (a,b) can be carried out in a FLL-theory. The set of the formulas of Sigma shows the different causes of the confusion of the characters a with the character b. For example, the two characters a and b may have the same sound (they are homophonic), or they may share the same graphic components, and so on.

In this paper, we first briefly outline the history of chinese characters [Alleton 1970], [Henshall 1988], [Li 1991] [Ryjick 1981], [Wieger 1978], so the reader can appreciate how a character is made up, how it acquired its structure and will make himself an opinion on the difficulties of the chinese writing. We also give the terminology we use. In the third section, we discuss the problem of characters representation and recognition which explains the limitation of our system. Then, after describing the system EMICW (section 4), we will give in section 5, an example of explanation in the FLL-theory T. The essential point of the section 6 is the proof of the tractability of our system.

2 Chinese Writing

The chinese characters originated between 3000- 2000 B.C in the Yellow River of China. They have been the subject of numerous studies. In this paper, we limit ourselves to mentioning what is essential for a good understanding of our work.

The chinese characters, also called **sinograms** (letters from China) are written in square form with the help of strokes, for example, horizontal stroke, vertical stroke. A set of 24 strokes standardized by the Foreign Languages Institute of Beijing are now of general use (see section 3.1). Strokes must be written down according to established principles of stroke order (generally from top to bottom, and from left to right) called **calligraphic order.** A knowledge of these principles is important in order to achieve the proper shape and to write in the cursive style or semi-cursive style (the writing style of the chineses). Sinograms are monosyllabic, and each syllable has a definite tone. There are **four basic tones** in the official national language (called mandarin chinese too). The transliteration used in this article is based on the official Chinese phonetic system, called **pinyin,** which is a representation of the sounds of the language in the Latin alphabet. We mark tones with numbers from 1 to 4. Sinograms have traditionally been classified into six categories. However, in many cases the categorization is

open to difference of opinion, and one sinogram can legitimately belong to more than one category. We list below the main categories that shed considerable light on the nature of sinograms. The students should consider these categories as guides to remembering sinograms.

1. The simple pictogram: essentially a picture of simple physical object. For example, woman 女 nu3, child 子 zi3.

2. The complex pictogram: a picture of several physical objects normally indissociable. For example, good 好 hao3.

3. The ideogram: a meaningful combination of two or more pictograms chosen for their meanings. For example, from pictograms sun 日 ri4, and moon 月 yue4, the ideogram intelligent is derived: 明

4. The ideo- phonogram: the largest category, containing about 90% of the sinograms. E ssentially a combination of a semantic element with a phonetic element. For example, the ideo-phonogram seed 籽 zi3 obtained by combining the semantic element cereal 米 mi3, with the phonetic element child 子 zi3, which gives to the character its reading. In fact, only about 30% of sinograms have a real phonetic component as in the example. Chinese (as any other language which is still spoken) has changed since the origin, so the phonetic element has lost its property.

The classification of sinograms in dictionaries can be done with the help of several methods. The number of strokes method and the alphabetical order (based on the pinyin romanization) method are easy to apply. The four corner method considers particular strokes located at the four corner of the sinogram. These strokes are codified with the help of four (or five) digits, and the sinogram is located at the position given by its numerical representation. The radical method uses a particular element in a sinogram, the key element, which indicates the general nature of the character. For instance, the ideo-phonogram 籽 zi3 is located under the radical 米 .

The character dictionary Xin Hua Zi Dian (eds.. 1979) lists the sinograms with respect to 189 radicals.

About five to seven thousand sinograms of up to ten or so strokes are needed in order to master the Chinese writing. The usual technique for learning consists in writing down a sinogram until it sinks in. We believe that the key to successful study of sinograms does not lie in rote learning. We propose a way to make the task a lot easier. For each case of mistake, our EMICW system gives an explanation based on the etymology of the characters. For instance, the character 天 tian1 (sky) can be confused with the following one 夫 fu4 (adult), because they have similar graphics. In fact, the character 天 comes from 大 da4 (tall), and from the graphic 一 yi1 (one), which represents a hat, while the character 夫 comes from 大 , and from the graphic ⺀ which means a hairpin. The position of the strokes can be meaningful. If such an explanation is given to the students in case of error, they progressively will be able to correct their own mistakes by reasoning, without relying heavily on memory. Moreover, they can consider these explanations as an introduction to the history of Eastern Asia.

We list below the main cases of mistakes we have met in our study of the chinese language:

1. **Confusion of homophonic sinograms:** about 50000 sinograms share four hundred syllables. According to official statistics each syllable with its tone corresponds to an average of five distinct sinograms. So, the first difficulty for students is to distinguish the homophonic sinograms.

For example, ten 十 shi2, moment 时 shi2, and to know 识 shi2 which are homophonic sinograms can be confused in a dictation.

2. **Confusion of sinograms with similar graphics:** For example, 己 ji3, 已 yi3, 巳 si4 have similar graphics, 天 tian1, 夫 fu4 adult have similar graphics too. It happens that the mistaken graphic is not a sinogram. For example, instead of half 半 ban1, the student (the author of this article) wrote 羊 .

3. **Confusion of sinograms which share the same components:** For example, 地 di4, and 池 chi2, which share the component 也 .

4. **Confusion of sinograms which form a word:** The sinograms are monosyllabic, but the chinese words are generally dissyllabic. For example, the words 身 体 shen1ti3 (body), 共 同 gong4tong2 (together), and 说 话 shuo1hua4 (to talk). The students usually learn dissyllabic words. So, they happen to confuse a sinogram with another.

We can also mention the case of confusion of simplified forms with non simplified forms of sinograms, of missing strokes: very complex sinograms may have about thirty strokes, so missing strokes is a very frequent mistake.

3 The Graphics Capture

Students write down sinograms from EMICW's dictation. A "good" method for representing graphics should allow the system to rapidly recognize the graphics drawn which are **not automatically sinograms**, because students can be mistaken. The different classification and search techniques in dictionaries that we have mentioned in the previous paragraph, permit to locate a character, but not to correct it. For instance, the four corners method does not take into account all the strokes drawn by the student, so, cannot be used to correct mistakes. The recognition problem of sinograms has been the subject of numerous studies. The last results can be found in [Wang 1988] [Yamamoto 1991].

3.1 Data Capture

In our particular application, we have to "understand" graphics drawn by students in order to help them in case of error. Each graphic drawn is characterized by the type of strokes used, the calligraphic order of strokes, and their positions in a square. In order to capture all these data, the system displays the set of 24 standardized strokes. In fact, only six strokes are primary ones: the point 丶 (pt), the horizontal stroke 一 (hr), the vertical stroke 丨 (vt), the top to left bottom stroke 丿 (dg), the top to right bottom stroke 乀 (dd), and the back up stroke 乁 (rt). All other strokes derived from these primary ones. These strokes are implemented by means of graphical primitives such as line drawing, rectangle and arc drawing. The students arrange strokes to draw graphics inside a square, the pictures may be expanded or shrunk to fit their destination square. For instance, the sinogram 天 tian1 (sky) can be written down in the following square by means of strokes of types hr, dg, and dd, according to the calligraphic order of writing (hr hr dg dd):

3.2 Graphic Feature of a Sinogram

As the position of strokes can be meaningful, we propose to locate each stroke in terms of coordinates on a plane (the coordinate plane is a two-dimension grid, which corresponds to the square drawn above, the coordinate origin (0, 0) being at the left top corner of the square). We sort out strokes with respect to their coordinates: from top to bottom (top-down order) , from bottom to top (bottom-up order), from left to right (left-right order), from right to left (right-left order). So, every graphic is characterized by the set of following codifications: the calligraphic order of strokes, the top-down, the bottom-up, the left-right and the right-left orders of strokes. For instance, the graphic feature of the sinogram 天 tian1 is given by the calligraphic order of strokes (hr hr dg dd), the top-bottom order (hr dg hr dd), the bottom-top order (dg dd hr hr), the left-right order (hr hr dg dd), and the right-left order (hr hr dd dg). We show now how all that knowledge can be used to explain graphic errors.

4 Knowledge Representation

The representation language of EMICW is a restricted version of the frame-based language KL-ONE [Brachman and Smolze 1985] - for instance it does not support structural dependency relations.

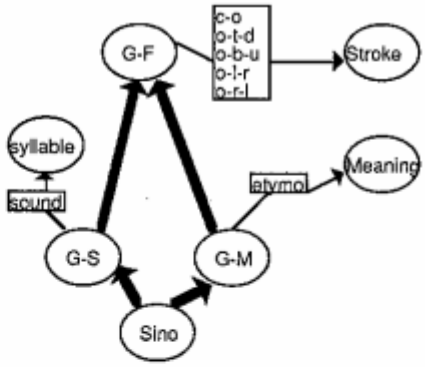
EMICW has a terminological component, the data base associated to an assertional component. The assertional component is a set of rules expressed in terms of predicates which are defined in the terminological component. Let us first justify our choice, then we will describe the language.

In order to deal with all the cases of mistakes listed in section 2, we need for a representation system which allows us to define all the links of "proximity" between the objects manipulated, i.e. graphics which are (or which are not) sinograms. For instance, homophonic links between two different sinograms, or graphic similarity between a graphic and its components. The inheritance link IS-A (B IS-A A means intuitively that all instances of B are also instances of A), and the properties which correspond to roles fit very well our problem. For efficiency reasons, we have to find a trade-off between the expressive power of the representation language and the computational tractability of the relation IS-A (called subsumption relation). In [Castaing 1991], we analysed the relation B IS-A A, and we proved that providing some restrictions, a subsumption criterion can be defined. A matching algorithm based on this criterion computes subsumption in polynomial time. In the system EMICW, we increase the expressive power of our language by adding to the system an assertional component, which only deals with existential rules. In section (6), we will discuss the computational complexity of our system.

4.1 Terminological component

Concepts are labelled collections of (attribute, value) pairs. The main concepts are the following ones:
Stroke, Graphic-Feature (abbreviated as G-F),
Graphic-Meaning (abbreviated as G-M), Graphic-Sound (abbreviated into G-S), Syllable, Meaning.
Individual concepts denoted by small letters are instances of concepts denoted by capital letters.
Attributes are classified into the structural link IS-A and properties.
The **IS-A link** is used for inheritance. So, if two concepts B and A are linked by means of the IS-A link, we say that A subsumes B, and that the concept B is of **type** A.
Properties are related to the intrinsic features of concepts. The attribute values are concepts too. The main properties in the system are the following ones: c-o (abbreviation of stroke calligraphic order), o-t-d (abbreviation of top-down order), o-b-u (abbreviation of bottom-up order), o-l-r (abbreviation of stroke order from left to right), o-r-l (abbreviation of stroke order from right to left), sound (pronunciation), etymo (abbreviation of etymology).

We give below a general view of the classification of the main concepts in EMICW taxonomy. To make clear the presentation, we use an ordering graph (semantic network), where the bold arrow -> represents the IS-A relation, and the arrow -> represents the roles.



In the taxonomy given above, there are only individual concepts of type Meaning. For instance, the words tall and hat are instances of Meaning. The concepts of type Syllable correspond to the syllables of the chinese language without tone. For instance, the concept Tian is of type Syllable. An instance of the concept Tian may be tian1 (first tone). The concepts of type Stroke correspond to ordered sequences of strokes. Let Sa and Sb be two concepts of type Stroke. Sb IS-A Sa if and only if the strokes in Sa also appear in Sb in the same order. For instance, the concept Sa which corresponds to the sequence of strokes (hr dg dd) subsumes the concept Sb given by the sequence (hr hr dg dd). Intuitively, this relation means that the graphics drawn by means of the ordered sequence of strokes (hr hr dg dd) have been partially drawn by means of the ordered sequence (hr dg dd) too. The concepts of type G-F give the graphic features of sinograms. The meaning of a sinogram

is given by the property etymo, and its reading is given by the property sound. It may happen that two different sinograms have the same graphic feature. For instance, to love 好 hao4, and good 好子 hao3. So, we define concepts of type G-M (Graphic Meaning), and G-S (Graphic Sound), such that each sinogram in the data base can be considered as an instance of the concepts G-M and G-S. We now give an example of a sinogram representation.

Example-1:
Let a100 be the sinogram 天 tian1. Its graphic feature can be defined by means of the concept G-F100 which is characterized by the following (attribute, value) pairs:
G-F100 = { (c-o, (hr, hr, dg, dd)), (o-t-d, (hr, dg, hr, dd)), (o-b-u, (dg, dd, hr, hr)), (o-l-r, (hr, hr, dg, dd)), (o-r-l, (hr, hr, dd, dg))}. The sinogram a100 inherits its meaning (sound) from the concept G-M100 (G-S100) partially defined by the following sets of (attribute, value) pairs:
G-M100 = {(IS-A G-F100), (etymo, sky) }
G-S100 = {(IS-A G-F100), (sound, tian)}
So, the sinogram a100 is an individual concept of type G-F100 defined by the (attribute, value) pairs: a100 ={(IS-A, G-F100), (etymo, sky), (sound, tian1) }.
End of the Example-1.

The graphics drawn by the student during a dictation are not automatically sinograms. So, we first consider them as concepts of type G-F (Graphic Feature). We solve the recognition problem of graphics by means of a **classifier** [Brachman and Levesque 1984].

4.1.1 Classifier

Usually the role of the classifier in a KL-ONE taxonomy consists in placing automatically a concept at its proper location. For classifying concepts in EMICW taxonomy, we proceed in two steps:
1. From the graphic drawn by the student, we define the concept CG (Complete- Graphic) related to the properties c-o, o-t-d, o-b-u, o-l-r, o-r-l of the components
2. We look for the concepts A and B, such that A subsumes CG, CG subsumes B, and there does not exist a concept A' which can be located between A and CG, and a concept B' which can be located between CG and B. We place CG, and we say that CG is at its **optimal location** in EMICW taxonomy. It means that CG inherits from all its ancestors. A is said to be a **father** of CG. B is said to be a **son** of CG. In case the concepts A and B are identical, we say that CG has been **identified** with A (or with B).

4.1.2 Recognition Problem

The recognition problem consists in discovering an individual concept b of type Sino, which has the same graphic feature than CG. We proceed as follows:
1. By means of the classifier, we place the concept CG at its optimal location.
2. If CG can be identified with a concept G-Fn of type G-F, it means that there exists at least a sinogram which is an instance of G-Mn and G-Sn. Let cf be this particular instance of G-Mn and G-Sn. We identify CG with cf, and CG "wins" all the properties of cf, for example, the properties sound, and etymo. We give an example.

Example-2
Let us suppose that the graphic drawn by the student is 天 tian1 (sky). The concept CG has the following properties (after sorting out the strokes with respect to their coordinates)
CG = { (c-o, (hr hr dg dd)), (o-t-d, (hr dg hr dd)), (o-b-u, (dg dd hr hr)), (o-l-r, (hr hr dg dd)), (o-r-l, (hr hr dd dg))}. The concept CG placed at its optimal location can be identified with the concept G-F100 (see the Example-1):
G-F100 = { (c-o, (hr hr dg dd)), (o-t-d, (hr dg hr dd)), (o-b-u, (dg dd hr hr)), (o-l-r, (hr hr dg dd)), (o-r-l, (hr hr dd dg))}, and so, can be identified with the instance a100 of G-M100 and G-S100. The concept CG gains the properties sound and etymo of a100.
End of the example-2.

Our recognition procedure is a little drastic. It may happen in sinograms with multiple components that some strokes in a component have no link with those in another component. By sorting out all strokes, we consider that they are necessarily linked, so, we detect a graphic error and reject the graphic proposed by the student. Our recognition procedure suits sinograms (simple or complex) whose components are specified by the students.

4.2 Rules

The rules of the assertional component deal with the different cases of error in chinese writing. All the predicates manipulated are defined in the terminological component either as unary predicates (concepts) or as binary predicates (roles), except for the predicates Error, ≠ (different), and = (equivalence). We explain now how the confusion of sinograms can be interpreted by means of the predicate Error.

Let a be the sinogram of the dictation, and CG be the complete concept obtained from the graphic drawn by the student. The student's answer is considered correct (there is no error) if and only if:
1. The concept CG is recognized as a sinogram denoted by b.
2. The individual concepts a and b share exactly the same properties.
Two cases of error are possible:
1. The concept CG cannot be identified with a concept of type Graphic- Feature of a sinogram. It means that the graphic drawn is not a sinogram.
2 The concept CG is recognized as a sinogram denoted by b, but the sinograms a and b do not share the same properties.

In the first case, the concept CG is located at its optimal position, and has a father that we denote by B. We consider an individual concept b of type B, and we propose to explain the confusion of a with b.The choice of an individual b may depend on a strategy. For the time being in our application, we identify CG with an individual which has the **same** graphic feature as B. In the second case, we propose to directly explain the confusion of a with b. The individual concepts pointed out by our system during an explanation are the **witnesses** of the error.
The rules of the assertional component have a limited syntax. Their general form is: "If there-exists x such that P (x) then Error (a, b)", where x is a vector of variables, and P is a finite conjunction of predicates. For instance, the

rule: "If there-exists z such that Syllable(z) & Sound(a, z) & Sound(b,z) then Error (a, b)", can be used in order to explain a mistake between two sinograms a and b which are homophonic. We give some examples of rules expressed in sequent calculus formalism.

rule-1: $\exists$ z Syllable(z) & Sound(a,z) & Sound(b, z) $\vdash$ Error (a, b)

rule-2: $\exists$ u z m1, m2, G-M(u) & G-M(z) & Meaning (m1) & Meaning (m2) & m1$\neq$m2 & u $\neq$ a & z $\neq$b & Etymo (u, m1) &Etymo(z, m2) & Etymo (a, m1) & Etymo (b, m2) & Error (u, z) $\vdash$ Error (a, b)

rule-3: $\exists$ s Stroke (s) & c-o(a, s) & c-o(b, s) $\vdash$ Error (a, b)

The rule-1 deals with errors due to homophonic sinograms. The rule-2 explains that the confusion of a with b may come from a misunderstanding of the etymologies of some components of the sinograms a and b.The rule-3 stresses the importance of the calligraphic order: two sinograms with the same calligraphic order can be confused.

5 Explanation in term of Proofs

In this section, we first present a formal description of EMICW by means of the FLL-theory T, then we will give an example of explanation. The FLL-theories use a fragment of linear logic (see also [Cerrito 1990], and [Masseron et al. 1990] for some particular applications of this logic). We suppose the reader familiar with sequent calculus. In the next chapter, we will discuss the tractability of EMICW.

5.1 Formal Description of EMICW

The FLL-theories are built from the linear fragment which consists of the connectives & (conjunction), the connective γ (disjunction), and the linear negation denoted by $(\ )^\circ$. The essential feature of the fragment used is the **absence of the contraction and weakening rules** listed below:

$$\frac{\Gamma, A, A \vdash \Delta}{\Gamma, A \vdash \Delta}\ (C\text{-}l) \qquad \frac{\Gamma \vdash \Delta, A, A}{\Gamma \vdash \Delta, A}\ (C\text{-}r)$$

$$\frac{\Gamma \vdash \Delta}{\Gamma, A \vdash \Delta}\ (W\text{-}l) \qquad \frac{\Gamma \vdash \Delta}{\Gamma \vdash \Delta, A}\ (W\text{-}r)$$

The axiom and the rules of the fragment are the following ones:

Axioms : $A \vdash A$

Cut : $$\frac{\Gamma \vdash \Delta, A \quad A, \Gamma' \vdash \Delta'}{\Gamma, \Gamma' \vdash \Delta, \Delta'}\ (C)$$

Exchange rules:

$$\frac{\Gamma, A, B \vdash \Delta}{\Gamma, B, A \vdash \Delta}\ (Ex\text{-}l) \qquad \frac{\Gamma \vdash \Delta, A, B}{\Gamma \vdash \Delta, B, A}\ (Ex\text{-}r)$$

Logical rules:

$$\frac{\Gamma \vdash A, \Delta}{\Gamma, A^\circ \vdash \Delta}\ (^\circ\text{-}l) \qquad \frac{\Gamma, A \vdash \Delta}{\Gamma \vdash A^\circ, \Delta}\ (^\circ\text{-}r)$$

$$\frac{\Gamma, A \vdash \Delta \quad \Gamma', B \vdash \Delta'}{\Gamma, \Gamma', (A\ \gamma\ B) \vdash \Delta, \Delta'}\ (\gamma\text{-}l)$$

$$\frac{\Gamma \vdash A, B, \Delta}{\Gamma \vdash (A\ \gamma\ B), \Delta}\ (\gamma\text{-}r)$$

$$\frac{\Gamma, A \vdash \Delta}{\Gamma, (A \,\&\, B) \vdash \Delta}\ (\&\text{-}l1) \qquad \frac{\Gamma, B \vdash \Delta}{\Gamma, (A \,\&\, B) \vdash \Delta}\ (\&\text{-}l2)$$

$$\frac{\Gamma \vdash A, \Delta \quad \Gamma \vdash B, \Delta}{\Gamma \vdash (A \& B), \Delta}\ (\&\text{-}r)$$

$$\frac{\Gamma, A(t/x) \vdash \Delta}{\Gamma, \forall x\ A \vdash \Delta}\ (\forall\text{-}l) \qquad \frac{\Gamma \vdash A, \Delta}{\Gamma \vdash \forall x\ A, \Delta}\ (\forall\text{-}r)$$

$$\frac{\Gamma, A \vdash \Delta}{\Gamma, \exists x\ A \vdash \Delta}\ (\exists\text{-}l) \qquad \frac{\Gamma \vdash A(t/x), \Delta}{\Gamma \vdash \exists x\ A, \Delta}\ (\exists\text{-}r)$$

In rules ($\forall$-r) and ($\exists$-l), x must not be free in Γ and Δ

A FLL-theory can be obtained from the above fragment by adding a finite set of proper axioms S, which are sequents closed under substitution. In the cut rule given above, the formula A is the **cut-formula**. A proof in a FLL-theory is said to be **cut-free**, if all cut-formulas involved occur in some sequent of S.

In our particular application, the set of proper axioms S which completely defines the FLL-theory T is made up of two subsets S1 and S2. The subset of proper axioms **S1** corresponds to the **terminological component**. They have the general form $A \vdash B$, where A and B are literals which interprete either concepts or roles. So, the terminological component of EMICW can be formally described by the FLL-theory T1 limited to the set of proper axioms S1. The subset of proper axioms **S2** is given by the rules of the **assertional component**.

5.2 To Explain is To Prove

EMICW combines the two following different reasoning methods:

1. The **classifier** which performs inferences by means of the subsumption operation.
2. A **theorem prover** which applies the cut-rule by only using the cut-formulas which appear in the rules of the set S2.

An explanation of a graphic error consists in finding a finite conjunction of **ground formulas** Sigma = P1 &...& Pn such that a proof of the linear sequent Sigma $\vdash$ Error (a, b) can be carried out in the FLL-theory T. Let us show how we proceed generally.

1. First case: the cut-formula doesn't contain the predicate Error.

$$\frac{\text{axiom of S2}}{\text{Sigma} \vdash \exists x\ P(x) \quad \exists x\ P(x) \vdash \text{Error}(a, b)}\ (\text{Cut})$$
$$\text{Sigma} \vdash \text{Error}(a, b)$$

The proof of the sequent Sigma $\vdash \exists x\ P(x)$ consists in instantiating the existential quantifier. We define a component called **instantiation component** which performs the following operations:
1. it defines a concept CP by using the properties given in the predicates P.
2. it locates the concept CP at its optimal position with the help of the classifier, such that there exists a witness c which satisfies P in the taxonomy of EMICW.

We obtain the new sequent to be proved, Sigma $\vdash$ P(c). We "force" the proof of this sequent by setting Sigma = P(c) & P2 ...&Pn. The proof of the sequent P(c) & P2 ...&Pn $\vdash$ P(c) is now straightforward by means of the (&-l1) rule.

2. Second case: the cut-formula contains the predicate Error. We are left with the following tree:

$$\frac{\text{Sigma} \vdash \exists x\ y\ P(x,y)\ \&\ \text{Error}(x, y)}{\text{Sigma} \vdash \text{Error}(a, b)}$$

In the same way as indicated above, we use the instantiation component to point out two witnesses c and d which satisfy P. We obtain the following sequent to be proved : Sigma $\vdash$ P(c,d) & Error (c,d). We apply the (&-r) rule and we obtain the new tree:

$$\frac{\text{Sigma} \vdash P(c,d) \qquad \text{Sigma} \vdash \text{Error}(c,d)}{\text{Sigma} \vdash P(c,d)\ \&\ \text{Error}(c,d).}\ (\&\text{-r})$$

We set Sigma = P(c,d) & P2...& Pn, so, we are now left with the proof of the sequent : P(c,d) & P2...& Pn $\vdash$ Error(c,d). We progressively makes appear all the formulas of Sigma by iterating the same process.

The sequent Sigma $\vdash$ Error (a,b) may have several proofs. In this case, the system can give multiple explanations to the students. The best explanation must allow the students to better memorize the sinogram a. We think that a good criterion for the choice of the best explanation can be :
1. The presence of the predicate Etymo in the explanation with the meanings of the components
2. The shorter proof (a proof which applies the smaller number of rules).

5.3 An Example of Explanation

Let us explain the confusion of the sinogram 天 tian1 (sky) with the sinogram 夫 fu4 (adult) by means of proofs. Etymologists give the following explainations: the sinogram sky comes from a person standing with arms spread out to look as **tall** as possible 大 with a big head (or a **hat**) symbolised by the stroke 一 . The sinogram adult comes from tall 大 with an ornamental **hairpin** through his hair (a sign of adulthood in ancient China) symbolised by the stroke 一 . So, we propose the following taxonomy:
1. The concepts G-F90 and G-M90 give the graphic feature of the sinogram 大 da4 (tall), and its etymology
G-F90 = { (c-o, (hr dg dd)), (o-t-d, (dg hr dd)), (o-b-u, (dg dd hr)), (o-l-r, (hr dg dd)), (o-r-l, (hr dd dg))}.
G-M90 = {(IS-A, G-F90), (etymo, tall)}.
2. The concept G-F01 corresponds to the graphic feature of the sinogram 一 yi1,
G-F01 = {(c-o, (hr)), (o-t-d, (hr)), (o-b-u, (hr)), (o-l-r, (hr)), (o-r-l, (hr))}. As the sinogram 一 yi1 has (at least) two different origins, hat and hairpin, we define two concepts of type G-M:
G-M010 = {(IS-A, G-F01), (etymo, hat)}
G-M011 = {(IS-A, G-F01), (etymo, hairpin)}
3. The concept G-M100 defined as {(IS-A, G-F100), (etymo, sky)} (see the example-1 of section 4.1) can be located now as:
G-M100 = {(IS-A , G-M90), (IS-A, G-M010)}.
The sinogram 天 tian1 represented by the individual concept a100= {(IS-A, G-M100), (IS-A, G-S100), (sound, tian1), (etymo, adult)} inherits the properties (etymo, tall) and (etymo, hat) from the concepts G-M90 and G-M010.
In the same way, the sinogram 夫 adult is represented by the individual concept b100= {(IS-A, G-M110), (IS-A, G-S110), (sound, fu4), (etymo, adult)}, and inherits the properties (etymo, tall), (etymo, hairpin) from the concepts G-M90 and G-M011.

In order to prove the sequent Sigma $\vdash$ Error (a100, b100), we propose to apply the cut-rule (C) with the cut-formula appearing in the rule-2:

$\exists$ u z m1 m2 G-M(u)& G-M(v) & Meaning (m1) & Meaning (m2) & m1 $\neq$ m2 & u $\neq$ a & z$\neq$ b & Etymo(u, m1) &Etymo(z, m2) & Etymo (a, m1) & Etymo (b, m2) & Error(u, z) $\vdash$ Error (a, b). We are left with the following sequent to be proved :

Sigma $\vdash \exists$ u z m1 m2 G-M(u) & G-M(v) & Meaning (m1) & Meaning (m2) & m1 $\neq$ m2 & u $\neq$ a100 & z$\neq$ b100 & Etymo(u, m1) & Etymo(z, m2) & Etymo (a100, m1) & Etymo (b100, m2) & Error(u, z).
The instantiation component instantiates the variable m1 to hat , and the variable m2 to hairpin (it has only this possibility), and defines the two individual concepts uC and vC whose etymologies correspond to these meanings: uC ={(IS-A, G-M010), (etymo, hat)}, and vC= {(IS-A, G-M011) , (etymo, hairpin)}.
Then, Sigma contains the following main ground formulas: Etymo(uC, hat) & Etymo(vC, hairpin) which shows that the reason of the confusion of a100 with b100 comes from a misunderstanding of the origins of the component 一 yi1 which appears in these two sinograms.
We invite the reader to try to apply the rule-3 in place of the rule-2. He will find that the confusion of a100 with b100 may come from the fact that these two sinograms have the same calligraphic order.

6 Computational Complexity

In this chapter, we prove that EMICW is tractable. The main problem comes from subsumption. The subsumption opeartion has been particularly analysed in [Levesque and

Brachman 1987] and in [Schmidt-Schaub 1989]. Their approach are mainly based on semantics. In [Castaing 1991], we have characterized a subsumption criterion by means of proofs in FLL-theories as T1 (see section 5.1). We briefly explain how we have proceeded.

6.1 Tractability of Subsumption

Let A and B be two concepts. We interpret A and B by means of first-order formulas, as in Brachman-Levesque's interpretation, then, we replace all classical connectives with linear ones. Let Ac = $\exists$xA1(x) &...An(x), and Bc = $\forall$zB1(z)&...Bm(z), (where z and x can be vectors of variables, and Ai(x) = $A_{i1}(x)$ γ...γ $A_{ip}(x)$, Bj(z) = $B_{j1}(z)$ γ...γ $B_{jq}(z)$) be the conjunctive normal forms obtained. A subsumes B iff there exists a cut-free proof in T1 of the sequent Bc $\vdash$ Ac. In the absence of contraction and weakening, we proved the following result :

Theorem (subsumption criterion): A subsumes B iff Ac and Bc satisfy the following **condition (C)**: there exists a, a substitution for x such that for each Ai, $1 \leq i \leq n$, there exists some Bj, $1 \leq j \leq m$, and b, a substitution for z, such that there exists a cut-free proof of the sequent Bj.b $\vdash$ Ai.a in the FLL-theory T1.

A matching algorithm can be easily derived from the condition C. It computes subsumption in polynomial time proportional to the length of the concepts, and to the cardinality of the set of proper axioms S1.

Without contraction and weakening, FLL-theories are decidable. There exists other decidable first-order theories which are based on classical logic [Ketonen and Weyhrauch 1984], or [Patel-Schneider 1985, 1988]. The originality of our approach comes from the way we deal with the universal quantifiers (or with the existential ones). Let us show how we can explain the rise in complexity of subsumption by means of contraction. We consider the following cases:

1. B_C and A_C satisfy condition (C) (the contraction rule is absent): the sequent $B_C \vdash A_C$ is provable in polynomial times, then the complexity of subsumption is **polynomial**.

2. Bc and Ac do not satisfy condition (C): let us suppose that the sequent $B_C \vdash A_C$ is provable (for example, by means of an approach based on semantics), and the proof of the sequent $B_C \vdash A_C$ necessitates the use of the contraction rule, (and possibly of the weakening rule): the search procedure for a proof can make sequents of the form $\forall z\, B(z,a) \vdash \Delta$, (or of the form $\Gamma \vdash \exists x\, A(x, a)$) appear at the nodes of the search-tree. Let us consider the case, where the sequent $\forall zB(z,a) \vdash \Delta$ appears at a node of the search-tree: the search procedure can go back-up the tree by applying the universal and contraction rules. We can be left with the following tree:

$$\dfrac{\dfrac{B\,(b\,/\,z,\,a),\ \forall z\ B(z,a)\ \vdash \Delta}{\forall z\ B(z,a),\ \forall z\ B(z,a) \vdash \Delta}(\forall\text{-l})}{\forall zB(z,a) \vdash \Delta}(\text{C-l})$$

The use of contraction may open a branch which terminates with a failure. Some back-tracking is then necessary. The complexity of the subsumption in this case is NP-hard.

3. $B_C \vdash A_C$ is not provable, then the use of the contraction rule may lead to duplicate infinitely the same formulas in the case where the set of instantiation terms (such as b) is infinite (for example in presence of functions) :

$$\dfrac{\dfrac{\dfrac{\vdots}{B(b/z,a),\forall zB(z,a),\forall zB(z,a)\vdash\Delta}(\text{C-l})}{B(b/z,\,a),\forall zB(z,a)\vdash\Delta}(\forall\text{-l})}{\forall z\ B(z,a),\forall z\ B(z,a) \vdash \Delta}(\text{C-l})$$
$$\forall zB(z,a) \vdash \Delta$$

The subsumption turns to be **undecidable.**

6.2 Tractabilty of EMICW

The terminological component of EMICW has a restricted syntax. The condition (C) defined above gives an adequate subsumption criterion. In order to locate a concept at its optimal location, the classifier performs the subsumption operations in number limited by the diameter of the semantic network. Its computational complexity is then limited. The theorem prover applies the cut-rule, with cut-formulas in some sequents of S2 (see section 5.2). Without contraction, the existential formulas which appear are **never duplicated**, and so, are only instantiated by means of the classifier. The cardinality of S2 is finite. Then, the proof of the sequent Sigma $\vdash$ Error (a,b) can also be carried out in limited time depending on the cardinaly of the set of proper axioms S= S1 + S2. The tractability of our system is then ensured.

Conclusion

A prototype of our EMICW system is implemented in LISP. For the time being, if the student writes down a graphic which is not recognised as a sinogram, the system has no particular strategy for discovering a "good" witness of the error. We are now investigating a strategy of choice of witnesses, which can take the context of the dictation, (the sinograms that the student have already drawn during the dictation) into account. Providing adequate rules, EMICW can also help students to learn japanese characters (kanjis) with the chinese or the japanese reading, or to learn classical vietnamese characters (nôm).

Acknowledgements

I would like to thank the four FGCS 's referees for their comments which contributed to clarify the presentation of my work. Discussions with my colleagues of PRC-IA were very helpful. The contribution of J-L. Lambert and C. Tollu to this work was invaluable. Thanks to both.

References

[Brachman R.J and Levesque H.J.1984]: "The Tractability of Subsumption in Frame-Based Descrition language"
Proceedings AAAI-84, August 84, pp34-37.

[Brachman R.J and Smolze J.G. 1985] : "An overwiew of the KL-ONE Knowledge Representation System". Cogitive Sci. 9(2) (1985) 171- 216.

[Brachman R.J and Gilbert V.P and Levesque H.J. 1985]: "An Essential Hybrid Reasoning System:Knowledge and Symbol Level Accounts of KRYPTON "
Proc. 9th IJCAI (1985) Los Angeles. pp 532-539.

[Bylander T. 1991]: "The Monotonic Abduction Problem: A Functional Characterization on the Edge Of Tractability".
Principles Of Knowledge Representation and reasoning Proceedings of the Second Internatipnal Conference. Cambridge, Massachusetts. April 1991.

[Cerrito S. 1990]: " A linear semantics for Allowed Logic Programs"
Proc. 5th Annual IEEE Symposium on Logic in Computer Science, IEEE Computer Society Press, 1990, 219-227.

[Castaing J. 1991]: "A New Formalisation Of Subsumption In Frame-Based Representation Systems".
Principles Of Knowledge Representation and Reasoning Proceedings of the Second International Conference. Cambridge, Massachusetts. April 1991.

[Girard J. Y. 1987] : "Linear Logic"
Theorical Computer Science 50 (1987) .pp1-102.

[Girard J. Y. 1989] :" Towards a Geometry of Interaction"
Proc. AMS Conference on Categories, Logic and Computer, Contemporary Mathematics 92, AMS 1989).

[Gentzen G.1969] :" The Collected Papers of Gerhard Gentzen"
Ed. E; Szabo, North-Holland, amsterdam (1969).

[Kazmareck T.S and Bates R and Robbins G.1986]: "Recent Developments in NIKL"
Proc. AAAI-86. Philadelphia, pp 978-985.

[Ketonen J. and Weyhrauch R. 1984] : "A decidable fragment of predicate calculus"
Theorical Computer Science 32:3, 1984.

[Levesque H.J and Brachman R.J. 1987]: "Expressiveness and Tractability in Knowledge Representation and Reasoning, Comp. Intell. 3 (2) (1987) pp 78-93.

[Masseron M. and Tollu C. and Vauzeilles J. 1990] : "Generating plans in linear logic"
Proc. FST & TCS 10, Bengalore (India), Dec. 1990.

[Nebel B. 1988]: "Computational Complexity of Terminological Reasoning in BACK".
Artificial Intelligence 34 (1988) pp371-383.

[Patel-Schneider P.F. (1985)] : "A Decidable First-Order Logic for Knowledge Representation"
Proceedings 9th. IJCAI (1985). Los Angeles. pp 455-458.

[Patel-Schneider P.F 1988]: "A Four-Valued Semantics for Terminological Logics"
Artificial Intelligence. 36 (1988) pp 319- 353.

[Schmidt-Schaub M. 1989]: "Subsumption in KL-ONE is Undecidable"
First International Conference on Principles of Knowledge Representation.1989. pp 421-431.

[Wang P.S.P 1988]: "On Line Chinese Character recognition"
6th IGC Int. Conference on Electronic Image pp209-214 1988.

[Yamamoto Y. 1991]: "Two-Dimensional Uniquely Parsable Isometrc Array Grammars".
Proceedings of the International Colloquium On Parallel Image Processing Paris June 1991.

[Alleton V.1970]: "L'Ecriture Chinoise".
Que sais-je N° 1374

[Bellassen 1989]: "Méthode d'Initiation à la Langue et à l'Ecriture chinoises".
Eds. La Compagnie/ Bellassen 1989.

[De Francis]: "Character Text for Beginning Chinese"
Yale Language Series. New haven and London, Yale University Press.

[Henshall Kenneth G 1988]. : "A Guide to Remembering Japanese Characters" Charles E. Tuttle Company, Inc. of Ruland, Vermont & Tokyo, Japan 1988.

[LI XiuQin 1991]: "Evolution de l'Ecriture Chinoise".
Librairie You Feng Paris1991.

[Lyssenko N. and Weulersse D.]: "Méthode Programmée du Chinois Moderne"
Eds. Lyssenko Paris 1987.

[Ryjick K. 1981] : "L'Idiot Chinois".
Payot 1981

[Shangai Foreign Language Institute 1982]: "A Concise Chinese Course For Foreign Learners" (Books 1 and 2).
Shangai Foreign Language Institute Press 1982.

[S.S Wieger 1978]: "Les caractères chinois"
Taichung 1978