Where jokers make the rules

Ralph Dessau, one of the developers of the low-price machine translation system PC-Translator, explains the use of "wildcards" in the system.

by
Ralph Dessau

Much has been said about artificial intelligence in machine translation, but one system uses a completely different approach. With PC-TRANSLATOR it is jokers or wildcards in its dictionary database which control the grammatical information.

Wildcards are marked as two question marks, **??**, followed by one or more letters. The first letter indicates word type, such as noun, adjective or verb. A second letter may specify gender, so that **??NMS** represents a noun in male gender, singular. In the dictionaries, the location of the translated wildcard is represented by the @ (arroba or "at" sign).

The dictionary files hold hundreds of wildcard phrases, as each phrase represents a grammatical rule. Critics have labelled this a simplistic approach, but it is actually quite the opposite. The highly effective coding of each dictionary entry helps wildcard phrases determine the gender of nouns and select the correct inflected form of adjectives. Other wildcard phrases conjugate verbs and select correct prepositions.

This concept has many advantages. All the dictionaries are ASCII files, which can be modified by word processing, and the coding of each entry is limited to a few letters. By contrast the dictionaries of many other MT systems are inaccessible to the user, and the coding of new entries requires operators with advanced degrees in linguistics.

Wildcards are extremely powerful, as each one may represent thousands of nouns, adjectives or regular verbs. Until recently, the concept was limited to one wildcard per phrase, but PC-TRANSLATOR can now accept multiple wildcards in a phrase. This dramatically increases its capacity for programmable grammatical information, and improves output quality accordingly.

A multiple wildcard phrase may look like this:

A ??AR ??NMS ,EN @0 @1,X,,

This phrase will correctly translate "A red car" in English to the Swedish "En röd bil", using the correct form of the indefinite article (En) and the adjective (röd).

Phrase entries are the keys to better output, as MT systems must learn a language word for word and phrase for phrase. Single words and ordinary phrases are used to define technical terminology or commercial idiom, wildcard phrases control the grammar of huge numbers of general expressions or sentences. Below follows a description of some important types of wildcard phrases.

Definite noun forms
(Only in Swedish and Danish)

NMD	("Real" Gender, Singular)
NMT	("Real" Gender, Plural)
NFD	("Neuter" Gender, Singular)
NFT	("Neuter" Gender, Plural)

The definite form of nouns in Swedish and Danish in some cases replaces the conventional form of a noun, preceded by the definite article **THE**. This difference from English grammar can be controlled by wildcard phrases such as:

THE ??AP ??NMD ,DEN @0 @1,X,,
THE ??AP ??NFD ,DET @0 @1,X,,

The above phrases enable PC-TRANSLATOR to translate "The white house" to "Det vita huset" and "The pretty girl" to "Den snygga flickan".

Wildcard phrases can also in some cases generate the definite form by adding the correct ending to the nouns. They look like this:

THE ??AP ??NMD ,DEN @0 @1EN,X,,
THE ??AP ??NFD ,DET @0 @1ET,X,,

Verb forms
Certain verb forms may carry indicators, which ensure their correct use in common grammatical constructions. Two such examples are the Gerund and the Past Participle, which look like this:

French	Gerund form:	**FINDING, TROUVANT,VG**
	Past Participle:	**FOUND,TROUVÉ,VP**

Tagging on a **G** or a **P** to the verb indicator **V** enables the wildcard phrases to deal correctly with some common grammatical rules.

The English gerund verb form is normally replaced in Swedish, Danish and Spanish by an infinitive. In Swedish, "By introducing" becomes "Genom att införa" and "By reading" becomes "Genom att läsa". Therefore the English gerund form "finding" is translated to Swedish as "att hitta". The dictionary entry looks like this:

FINDING,ATT HITTA,VG

while the corresponding wildcard phrase looks like this:

BY ??VG ,GENOM @,X,,

A similar rule exists in Spanish. "By introducing" becomes "Al introducir" and "By reading" becomes "Al leer". The wildcard phrase looks like this:

BY ??VG ,AL @,X,,

In contrast, French uses the gerund verb form, preceded by the word "En", so that "By introducing" becomes "En *introduisant*" and "By *reading*" becomes "En lisant". The corresponding French wildcard phrase therefore looks like this:

BY ??VG ,EN @,X,,

In other instances the English gerund is replaced by the imperfect form of the same verb in French. "He was sleeping" becomes "Il dormait" and "She was writing" becomes "Elle écrivait". "She was not writing", however, in French becomes "Elle n'écrivait pas". To cope with the contraction from "ne" to "n'", it was necessary to add a verb form, which not only replaces a gerund with imperfect, but also indicates if the first letter of the French verb is a vowel or a consonant. It looks like this:

WRITING,ÉCRIVAIT,VGV
SLEEPING,DORMAIT,VGC

Wildcard phrases controlling these verbs are therefore entered in the dictionary like this:

WAS ??VGV ,@,X,,
WAS ??VGC ,@,X,,
WAS NOT ??VGV ,N'@ PAS,X,,
WAS NOT ??VGC ,NE @ PAS,X,,
WASN'T ??VGV ,N'@ PAS,X,,
WASN'T ??VGC ,NE @ PAS,X,,

Similar grammatical rules also apply to the past participle verb form in certain languages, such as Swedish, Danish and French, and represent the reason for coding this verb form with a **P**.

Certain French and Italian verb forms have indicators to show when the first letter is a vowel or a consonant. The indicator tells the program when to replace the "e" in "de", "je" and "ne" with an apostrophe:

de trouver	(to find)
d'écouter	(to hear)
de ne trouver	(not to find)
de n'écouter	(not to hear)
Je trouve	(I find)
J'écoute	(I hear)

Two of the verb forms which use vowel/consonant indicators are:

Verb Infinitive: **HEAR,ÉCOUTER,VIV**
FIND,TROUVER,VIC
Verbs Pres. Ind., 3rd Pers., Sing.:
HEARS,ÉCOUTE,V3V
FINDS,TROUVE,V3C

Verb conjugation

Regular verb conjugation by wildcards is made possible by adding their roots in the dictionaries. Verb-root wildcards use a **W** to distinguish them from actual verbs, while a digit indicates which conjugation they follow. In French and Italian, a third indicator may show whether the first letter of the verb is a consonant or a vowel. A typical verb-root wildcard may be **W1C** for a first conjugation verb starting with a consonant. Examples of wildcard phrases, conjugating French verbs from the English equivalent with auxiliary verbs, may look like this:

HE WILL ??W1C ,IL @ERA,X,,
HE WILL NOT ??W1V ME ,IL NE M'@ERA PAS,X,,

The first phrase will correctly translate "He will forget" to French "Il oubliera". The second phrase will translate "He will not forget me" to "Il ne m'oubliera pas".

The coding is extremely flexible and can easily be made to include sub-forms of verbs within the major conjugations or other word categories.

Applications

In actual practice, the program works with sequences of these phrases. For example, "He became ill by drinking the contaminated water" would use a set of wildcard phrases which might look like this:

HE BECAME ??AMS ,IL DEVENAIT @,X,,
BY ??VG ,EN @,X,,
THE ??AFS ??NFSV ,L'@1 @0,X,,

The output would read:

"Il devenait malade en buvant l'eau contaminée".

The same set of phrases would translate:

"He became angry by hearing the unfavourable opinion"

to:

"Il devenait fâché en écoutant l'opinion défavorable".

At this point, innocent readers might ask: why go to all this trouble to translate simple stuff like this? If so, they may never have seen a software instruction manual or a truck repair handbook, both of which are packed with statements such as:

Select Setup from the File Menu
Select Minutes between Backup
Press Exit to exit the Setup Menu

or

Remove the circlip
Press off the guide sleeve
Remove the engaging sleeve

A software manual may easily contain more than 200 pages, while a truck manual might exceed 500. Both software and trucks are exported, and must be accompanied by documentation in the local language of the importing country.

Suddenly the jokers begin to have a useful purpose. Even more so, when the initial version of the software is followed by others, and new truck models keep entering the market at regular intervals.

In these applications, the intelligibility of a translation cannot be a matter of judgement. Statements such as "Remove the circlip" or "Press off the guide sleeve" must be translated exactly and without ambiguity. The wildcard technique is ideal for such applications.

Translations are made in a wide range of subjects and writing styles. A major part, however, involves repetitive but highly specific technical and commercial documentation. Because of their special terminologies this is a costly category, and is therefore also the area where MT is most cost-effective.

When PC-TRANSLATOR is used in this type of translation, it can be made to safeguard existing vocabularies and to reproduce established output styles. As human translators come and go, a good system should preserve and enhance their contributions. This is important, because writers in different corporate cultures or professional environments write about identical topics in a wide variety of styles.

The PC-TRANSLATOR technique of using wildcard and conventional phrases is well suited for style adaptation. It can, for example, be made to translate conventional language in a source text to legalese and vice versa. The advantages become even more pronounced when PC-TRANSLATOR is employed on large volumes of text.

Conclusion

The MT system described herein has proved highly effective. Because of its modular structure, new language pairs can be added much faster than for any other system. The transparency of its dictionaries makes them easy to modify, and the wildcard phrases offer an extremely precise and logical mechanism to define grammatical rules. Specific terminology and idiom can be dealt with just as effectively with conventional phrases.

PC-TRANSLATOR was the first PC-based MT system, at its introduction in 1985. Then as now, an increase in dictionary entries reduces the speed. PCs have since improved enormously, with faster CPUs and more memory, but memory and speed are still the main obstacles to better output quality.

Ideally all grammatical rules should be included, but this would reduce the speed to zero. Instead, enhancements follow technological advances. Thus the faster CPUs and lower-cost memory devices which are currently on the drawing board will soon allow us to add more wildcard phrases and thereby achieve an even higher output quality.

Although most systems currently translate to or from English, they can perform even better when used between related languages, such as French, Italian, Spanish and Portuguese. Linguistic Products, the company behind PC-TRANSLATOR, will be able to generate the dictionaries for such language pairs by computer because of its modular wildcard architecture.

In summary, we can conclude that the wildcard phrase approach to MT is both viable and effective. Linguistic aspects can be represented as thoroughly as by any other known process and in a much simpler manner. The modular structure of these systems further facilitates the creation and eventual use of additional language pairs.