Abstract: We define non-expected language as the set of words and phrases which do not match exactly the syntactic, lexical and conversational expectations of a natural language processing system.

We consider three different kinds of mistakes and mismatches:
- lexical mistakes: misspelt or unknown words,
- syntactic mistakes: phrases which are either misconstructed or not described by standard grammar,
- semantic or pragmatic mistakes: sentences which are either difficultly understood or in contradiction with the system's knowledge.

This processing of non-expected language will be applied to intelligent word processing and conversational access to knowledge base.

1. Introduction

The understanding mechanisms may fail at different steps because of various reasons: at the lexical level, when there appears an unknown word or a word whose sense does not match with the context; at the syntactic level, when a sentence cannot be interpreted; at the discourse level, when a sentence is in conflict with the previous ones. We present here our approach to solve these problems.

2. Lexical level

The usual spelling correction methods use special purpose dictionaries. The chosen structure implies often which results can be obtained: an alphabetical sorted dictionary will obtain very speedy corrections for words with the well-spelled beginning, and probably bad ones if the fault place is at the beginning (for example: magravation instead of aggravation).

The spelling correction methods which use similarity keys (for example: the speedcop correction algorithm [Pollack 84]) work also on special dictionaries: they are sorted following the keys, and cannot be used without them.

When the spelling correction is the only goal we want to achieve, this particularity makes no trouble. In a natural language understanding system, we must deal with syntactic and semantic information. There are treatments which look for that information, and have to find them very quickly. A spelling correction sorted dictionary is probably a very bad structure for semantic oriented searches: it is able to tell you that totally looks like totally, but not that thoroughly refers to the same as totally.

It seems to us, that the internal organization of the dictionary and the access methods to use it must be dissociated. Our access method introduces an intermediary level between the searched words and a non-specifically structured dictionary.

This level consists in a B-tree-like structure. We use a spelling oriented similarity key, the anacode [Fournier 86], which has the particularity of being insensitive to the order of the letters in a given word. These anacodes are the pivots with which we build the non-terminal nodes of the tree. The terminal nodes refer to the entry points of the great dictionary. Other intermediate levels can easily be built with different goals, for example conceptual analogies, and with different access methods.

This architecture allows the multiplicity of such intermediary levels (for example a typographically oriented one and a phonetically oriented one), and an easy introduction of parallelism. An independent intermediate structure allows distributed working, even on specialized systems.
3. Syntactic level

The mistakes come from unknown constructions: repetitions, omissions, interventions and substitutions of words, or from bad agreements between constituents. This occurs between the subject and the verb, the determinant and the noun, the adjective and the noun.

There are different approaches to solve these problems. One consists in building a grammar of errors. This can be viewed as the a priori description of variant language use. One finds immediately that it is very difficult to achieve exhaustivity. This will be discussed in section 4.

Another approach [Carbonell 83] bypasses the syntactic problems doing a purely semantic analysis. However it is possible only in a well-known domain.

A third method consists in designing a parser which can operate an analysis of well-formed groups (islands of confidence) and make binding hypotheses to obtain a correct interpretation of the sentence [Fouqueré 88].

Our work [Galiacy 88], is independent of any domain and must be used by a wide audience; we find it excessive to write a new parser accommodating language variation. Therefore we prefer to augment an existing parser by implementing a specific module that runs when the parser fails.

Before describing this module, we briefly present the parser [Francopoulo 88]. It operates by chunking: it first builds nominal, verbal and adjectival groups in isolation. Then it tries to connect them according to syntactic and semantic rules.

Mistakes are detected when the parser cannot interpret a sentence because it is ungrammatical, although each word is recognized. At this stage, the constituents of the sentence are available for further processing.

We apply three strategies to interpret erroneous sentences, according to the type of the error:

- when two words of a noun phrase or a subject and a verb do not agree, we can identify the rule which cannot apply. Then, the module inhibits this rule and parsing can be resumed. The internal representation is then corrected by applying phonetical, lexical and psychological criteria.

Let us illustrate this strategy by some examples:

a) ce chevaux (this horses) ---＞ ces chevaux (these horses)

b) un cheval (a horse) rather than des chevaux (horses)

c) ce chats (this cats) ---＞ ces chats (these cats) rather than ce chat (this cat)

- In the first example, all criteria point to the same solution; in the second, using only phonetical and lexical criteria, un is preferred over des, and chevaux over cheval, which means a conflict. Fortunately, the psychological criterion says that articles are known better than irregular plurals, which makes the choice possible. In the last, the situation is similar: the first two criteria reach no decision, and again a psychological one applies. It is a principle of least effort (a letter is omitted rather than added). In the future, psychological criteria will be attached to the user model (cf part 4).

- In case of redundant noun phrases, we correct the semantic structure: there are two candidates for one case role and they refer to the same thing; so only one is kept.

La pomme, je la mange (the apple, I eat it)

- There are many other cases of errors: omission of functional words, the use of a preposition for another and too many verb groups. The kind of problem is detected by specialized rules which take the appropriate action on the surface structure. For example,

Vous avez trier les questions (You have sort the questions)

is seen as having a verb in excess; the action consists in converting the verb into participial form.

While our approach does not cover all kind of sentences which make parsing difficult, it relies upon general principles and so can be extended in a systematic way.

4. The conversational level

We have so far presented treatments for deviant spellings or agreements. Are there any other kinds of communication problems? In principle, a sentence consisting of acceptable words in a correct order and with the right agreements could never cause any trouble in understanding... except if it does not fit in the current conversation (or linguistic exchange, more generally speaking). Among the various examples one could think of, let us select one [Suchard 87, question 163]:

EUROSPEECH '89, Paris, France, September 1989
Pour permettre au jeunes de rester en formation après la fin d'étude [for études]

(To enable young people to continue training after they finish their study [for studies])

Here is a situation where a morpho-semantic confusion (singular/plural with different meanings) occurs in a semi-alignable conflict ("to finish a study" is not like "finishing one's studies"). Hence we put it in the class of "morpho-pragmatical problems".

Keeping this particularity in mind, we develop the main points of our strategy:

a) how morpho-pragmatical problems are detected
b) how to accomodate them (focussing on the morpho-semantic part)
c) how to go on with the conversation when b) fails
d) how to sum up the conflict in order to properly terminate the interaction if c) itself would fail.

There are two ways of detecting a morpho-pragmatical problem: either a selectional restriction is violated, or a "surprising topic change" occurs. By the latter, we mean that the "conversational continuity" [Vilnat 84] is not respected. In our example, the exchange is about people looking for jobs; so the notion of "finishing a study" appears indeed unrelated to the general topic.

At this stage, the system looks for morpho-semantic rules which are easily violated. In order to represent "which rule is respected by whom", we maintain a two-dimensional table (inspired by [Kilbury 86]) whose lines correspond to rules, the columns being attributed to user models. A rule is "easily violated" if the corresponding line contains many negative indications, meaning "this rule does not apply to the discourse of this user". So we can express the fact that some users do not reliably indicate if they have written a noun in the singular or plural: for such users, one cannot apply rules of the form "this word means X when singular and Y when plural". Fortunately, there may exist other rules discriminating between X and Y, as, in our example, the adjective "professionnelles" which matches with "études" rather than "étude".

If, however, the user's confusion of two words goes too far, our system may be unable to find such discriminating rules. In that case, the difficulty is signalled to the user by a message like "The distinction between X and Y is an essential one, and since it does not appear clearly for me in your sentences, we cannot deal with these notions". This does not prevent the linguistic exchange to go on, leaving aside any direct mention of the troublesome topics.

An extreme case would be a situation where such communication failures would occur repeatedly: this would mean that the whole attitude of the user is in conflict with the one embodied by the system. The exchange should then come to a conclusion. The last task of the system being to build a description of the semantic discrepancies which lead to the communication failure.

To summarize our approach, we could say that we insist on designing conversational systems that, once they have "intelligently" detected a conflict, show no less intelligence in continuing the conversation within the altered conditions. This is achieved by informing the user of the difficulties his language use creates in the system's functioning.

5. Conclusion

We want to emphasize the generality of our approach: the lexical treatment does not require a specialized dictionary, the other treatments are independent of the domain and can be used by a wide audience.

We intend to integrate these three separate aspects in a general model of natural language understanding (CHAMEL). This model is able to perform various tasks, by adapting its behaviour. Thus our work will have to be extended, in order to interpret, suggest corrections, or even correct, according to the specific task: text processing assistance, intelligent tutorial system, text understanding, question-answering system, ...

6. Bibliography

Vilnat A. 1984 L'élaboration de descriptions des relations conceptuelles, Thèse de l' Université Paris Nord

Kilbury J. 1986 Correction automatique dans les systèmes question-réponse en langage naturel, Thèse de l' Université Paris Nord

Fournier J. P. 1986 Correction automatique dans les systèmes question-réponse en langage naturel, Thèse de l' Université Paris Nord

Berthelot J. 1. 1989 Analyse contextuelle et régulation interprétative, Thèse de l' Université Paris Nord

Pollock J. 1. and Louisa A. 1984 Automatic spelling correction in scientific and scholarly texts, Communications ACM 27, 4, pp 358-368

5. Conclusion

An extreme case would be a situation where such communication failures would occur repeatedly: this would mean that the whole attitude of the user is in conflict with the one embodied by the system. The exchange should then come to a conclusion. The last task of the system being to build a description of the semantic discrepancies which lead to the communication failure.

To summarize our approach, we could say that we insist on designing conversational systems that, once they have "intelligently" detected a conflict, show no less intelligence in continuing the conversation within the altered conditions. This is achieved by informing the user of the difficulties his language use creates in the system's functioning.

5. Conclusion

We want to emphasize the generality of our approach: the lexical treatment does not require a specialized dictionary, the other treatments are independent of the domain and can be used by a wide audience.

We intend to integrate these three separate aspects in a general model of natural language understanding (CHAMEL). This model is able to perform various tasks, by adapting its behaviour. Thus our work will have to be extended, in order to interpret, suggest corrections, or even correct, according to the specific task: text processing assistance, intelligent tutorial system, text understanding, question-answering system, ...

6. Bibliography

Vilnat A. 1984 L'établissement des interventions pertinentes dans une conversation homme-machine, Thèse de l' Université Paris Nord