MENTOR88/P: A BRIEF OVERVIEW

February 28th, 1989

Diana Santos

IBM-INESC Scientific Group
R. Alves Redol, 9 - 4 Dto
1000 Lisboa
Portugal
# Table of Contents

Introduction ................................................................. 1  
The system ................................................................. 1  
Functional view of MENTOR88/P architecture .................... 1  
Our approach to structural transfer ................................ 2  
Our approach to tense transfer ........................................ 3  
Our approach to lexical transfer ...................................... 3  
Our approach to generation ............................................ 4  
Dealing with MWEs in MENTOR88/P .................................. 4  
Evaluation of the system ................................................ 6  
Conclusions ..................................................................... 7  
References ........................................................................ 7  
Sample translations .......................................................... 7
Introduction

This small text is intended to give a panoramic view of the achievements and current state of the Portuguese machine translation prototype that is being developed in the IBM-INESC Scientific Group, in the framework of the MENTOR project.

The system

The most immediate feature that differentiates MENTOR88/P[7] from the approaches taken in the other MENTOR sites is that it was all built in its entirety in the framework of PLNLP[1]. In fact, following the example of the English parser MENTOR is using, PEG[2], it was decided to make use of PLNLP as the environment for natural language processing.

At this moment in Lisbon we have a single environment that can run in two modes: either accepting English text to be analysed by PEG and afterwards translated by MENTOR88/P, or allowing Portuguese text to be input and analysed by a new PLNLP grammar for Portuguese, whose development started in November 88.

This integration of two NLP programs is beneficial from several points of view, ranging from the simple power of sharing a common environment and tools, to the fact that the two programs (the translator and the Portuguese parser) are actually related in several other ways.

In fact, one of the main goals of MENTOR88/P from the beginning was to produce a translation system that was modular, in the sense that the output of the transfer phase should be a description of the Portuguese sentence, independent of the process that built it. In other words, we require the input to the generation phase of the translation system to be an object (a PLNLP graph) that has only to do with Portuguese, thus keeping no trace from English. It is easy to see that this structural description is exactly also the aim of the PLNLP grammar for Portuguese. Therefore, the existence of this last grammar provides us with a way to check the correctness and coherence of our transfer (translation) rules. Moreover, in “Dealing with MWEs in MENTOR88/P” on page 4 it will be shown that this grammar actually also takes part in the translation process.

Functional view of MENTOR88/P architecture

Let us describe with a little detail the structure of MENTOR88/P, in terms of the functions performed by the system.

By looking at Figure 1, several important features can be pointed out:

- The transfer phase is the bridge between two independent internal structures, one for English, created by PEG, and one for Portuguese, the input for generation phase (and output of transfer).
- We distinguish between three kinds of basic transfer
  
  structural relating the different structures of the two languages
  lexical relating the different words or expressions of the two languages
  tense relating the two different tense systems

- Basic transfer is, for us, the selection of all possible alternatives that make sense as translations into Portuguese. Style transfer is how to decide on only one. We have not begun working on this last problem yet, so we select arbitrarily the first.

- The three kinds of basic transfer are evidently related among them, and our design accommodates this fact. So, the choice of a particular lexical item can ask for a specific structure, as well as the choice of a particular structure (or word) can imply a given tense. On the other hand, it is based on its structural (and aspectual) context that the lexical choice of a word is performed.

- Finally, there are different conceptual levels at which all the modules in the diagram can be viewed, related to the knowledge they use. At the moment, it is mainly at the syntactical (surface and deep) levels that work has been produced, but we intend to use discourse knowledge and other levels of language description very soon.
We proceed now to give a more detailed view of the three basic modules of transfer, and of generation.

Our approach to structural transfer

PEG furnishes a very detailed account of the structure of the input sentence, with a graph that both mirrors the surface structure and provides deep syntactical features, like object / subject. That graph comprehends also information about the subcategorization of the particular words involved, together with a classification of the particular structure in question, and information about aspect and tense features.

It was our goal to optimize the use of all the features and information provided, in a modular and organized way. In MENTOR88/P, every node in the graph is visited only once in a top-down left-to-right manner, and is submitted to all rules that can apply to its structure.

Structural rules are written in such a way that each one (or a group of them) solves a particular problem, not interfering with the others, thus constituting a modular, readable and easily extendable piece of code.

Moreover, and due to the concept of "computed structure" [3] that makes each node of the graph independent of the way it was computed during parsing and therefore carrying a definite linguistic meaning, rules were devised to operate at the level of a single node. That is, a structure rule in MENTOR88/P can only alter the node it applies to, and not an arbitrary portion of the graph. This does not mean, obviously, that complex structure changes do not take place, but only that this process is done through an ordered treatment of the whole graph.

As a by-product of our work on structural transfer, we developed a number of PLNLP procedures that form the basic actions and tests needed for the task, defining thus a specialized language.

Finally, it is important to emphasize, too, the influence that lexical transfer has in the structure computed. As can be seen in the section below devoted to lexical transfer, and most particularly in Figure 2 on page 3, the choice of a particular Portuguese word can impose a particular structure.
Our approach to tense transfer

The different verbal systems of the two languages - with a considerably bigger choice among tenses in Portuguese - requires a detailed contrastive study of the tense meanings, together with the consideration of aspect, and a temporal categorization of adverbial and prepositional phrases.

Our work on this area has produced a set of "tense transfer" rules that handle English present and past clauses and produce their translation into the corresponding several Portuguese tense equivalents. We make use of such diversified information as relationships among clauses, kind of adverbial modification, and the particular Portuguese verb selected.

Our approach to lexical transfer

We interleave structural and lexical transfer, so that only after (and as soon as) all nodes immediately below\(^1\) one particular node get lexically transferred, are the structural rules applied to that node. Then the process starts again with lexical and structural processing, in a top-down traversal.

Concerning lexical transfer (LT) proper,

- As opposed to perform LT only at the graph's leaves, the above mentioned interleaving results in performing lexical transfer as soon as we find the lexical item in the PEG graph, which corresponds to access the bilingual dictionary with the (lemma of the) head word of each node.
- For the cases where more than one Portuguese lexical item corresponds to an English word, (a very frequent situation, actually), the conditions to select between the different alternatives have to be stated (as in any MT system). We chose to associate the conditions under which one word is a valid translation to the Portuguese (as opposed to the English) word.

\[
\begin{align*}
\text{acautelar}(\text{VERB} & \text{ (NCOND self NEG) (COND SENTRYPE IMPR)} \\
\text{REFLEX} & \text{ (PREPO OBJECT com)})
\end{align*}
\]

\[
\begin{align*}
\text{relógio}(\text{NOUN}) \\
\text{observar}(\text{VERB} & \text{ (COND self PASSIVE)}) \\
\text{observar}(\text{VERB} & \text{ (COND self PRESPART)}) \\
\text{ver}(\text{VERB} & \text{ (COND self NEG)}) \\
\text{ver}(\text{VERB} & \text{ (COND self OBJECT)}) \\
\text{cuidar}(\text{VERB} & \text{ (NCOND self PASSIVE self NEG) (COND self OBJECT OBJECT ANIM)})
\end{align*}
\]

Figure 2. Lexical transfer of "watch": Four different verbs are contemplated as possible translations for the verb "to watch". NCOND and COND indicate respectively exclusive and necessary conditions. REFLEX and PREPO are structural features of concern to structural transfer.

This has as consequence that each translation is independently stated, and that the system allows for more\(^2\) than one translation to be selected. This increases, in our opinion, both the readability and the ease of the building of the dictionaries, and has the pleasant feature of giving the same treatment to structural and disambiguation (LT) information, in that they are both associated with the target language item.

- As for the knowledge actually used to state the conditions above mentioned, we make use of the context in which the words occur, using two kinds of different information, both available in PEG's parse trees: the structure in which those words take part, and the subcategorization in terms of "universal features"\(^3\) of the words related, normally through functional concepts (like subject or object).

\(^1\) In a phrase structure grammar, "below" stands for "dominated by".

\(^2\) Actually, it also allows for NO translation at all in case the context in question had not been identified. This constitutes, in our opinion, in a development phase, a preferable alternative to a default (and therefore non-monotonic) system.

\(^3\) By universal features we mean markers like ANImed, HUMan or TiME that represent general characteristics independent of domain or language.
Another distinction can be drawn between two kinds of context we use - the particularities of the dominated constituents, and the general type of the domination structure. While this last amounts to propagate general indications like NEGation, TiME, or QUESTION downwards, the first kind of information must be searched through the specific attributes of the current node.

It has been our concern to detect general patterns applicable to lexical transfer as a whole, aiming at generalization and identification of which features are (more) relevant.

**Our approach to generation**

Generation in MENTOR88/P has to do with the creation of the ultimate output string that is the Portuguese sentence. As we start from a Portuguese graph structure, it is important to emphasize that this last is considered to have no traces from English or whatever language it came from.

Therefore, generation tasks are only concerned with the creation of the best surface string that can be associated to that graph. Depending on the level at which the features of interest are represented, (which in turn depends on the level where transfer is performed), more or less work is required to produce the desired output string. The only uncontroversial task is the morphological processing that has to be done as the last stage to produce a natural sentence.

But even for the inflection and conjugation of the words another more delicate process has to take place, namely, the traversal of the graph. It is necessary to define, based on several linguistic rules, the dominating morphology features (like gender, number, or case for pronouns) that will have then to be propagated with care, in order to produce the agreement network required by the Portuguese language.

As examples of other generation tasks, we should mention the delicate problem of the ellision of subject pronouns, and the positioning of personal pronouns in a clause, or of adjectives in a noun phrase. Many target language factors, like the length of modifiers, the danger of ambiguity or simply statistical preferences can influence the result.

Finally, another problem to handle before a true sentence is born is the contraction and hyphenation that are characteristic of Portuguese, and that cannot be done as a mere string post-processing phase, since they require knowledge about syntactic and even deep syntactic information.⁴

**Dealing with MWEs in MENTOR88/P**

The fact that we have access to the very powerful PLNLP environment described in the beginning of this text, namely with grammars of the two languages involved available, allowed us to sketch a different solution for dealing with the multiword expression problem in translation.

It is a well known fact that there is not a one-to-one correspondence between the words of two different languages, and one consequence of this is the frequent translation of one English word, for instance, into an (arbitrarily complex) Portuguese expression, a multiword expression (MWE). At this moment, we are thus concerned with the translation of cases when a single English word is enough to describe a concept, for which Portuguese speakers use a combination of words, like "to miss", "sentir a falta".⁵

As far as we know, this problem has been taken into consideration in the machine translation world by storing in the bilingual dictionary the structure (in the representation the system is using) of the

---

⁴ For instance, the preposition "de" in "de o" should be contracted into "do" if and only if "o" is a definite article, and never if it is a personal pronoun; the verb followed by the indirect object personal pronoun "lhe" "deu lhe a" should only be produced as "deu-lha" if "a" is the direct object pronoun, while it should remain "deu-lhe a" if "a" is an article.

⁵ It is important to emphasize that by MWE we do not mean a sequence of characters that just by chance has some spaces in it (like "according to") but the expression of a concept by several words whose grammatical function is not altered, like "have fun" or "ride a bicycle" or "be in love". An important characteristic of these expressions, moreover, is that the words do not need to be contiguous in the sentence.

⁶ This is the opposite situation to the examples mentioned in⁵.
phrase in the target language, which will then be introduced in the place of the source word. (See [8] or [6]).

Our approach was to store in the bilingual dictionary the actual Portuguese (target language) string (as in a human dictionary), and parse it with our Portuguese grammar, when that particular translation is required.

This provides minimal storage in the bilingual dictionary together with maximal readability. It conveys an appreciable power, too, in that the result of the parse of the MWE can be then dealt with, by the generation process, making use of all general rules embedded in the system.

The following example may help to clarify this point:

\[
\begin{align*}
\text{sentir(VERB CHTPOSS (EVP sentir a falta) (COND \\
\text{SUBJECT ANIM self OBJECT}))} \\
\text{sentir(VERB CHTPOSS (EVP sentir a falta) (COND self PASSIVE))} \\
\text{perder(VERB (NCOND OBJECT ANIM) (COND SUBJECT ANIM self OBJECT))} \\
\text{faltar(VERB (NCOND self OBJECT self PASSIVE) (COND self PROG))} \\
\text{deixar(VERB (EVP deixar_ escapar) (COND self OBJECT))} \\
\text{menina(NOUN)}
\end{align*}
\]

Figure 3. Example of lexical transfer of the word “miss” into a MWE: EVP stands for MWE, and the value of this attribute is precisely the multiword expression that should replace the single English verb “miss”.

together with the presentation of the actual output produced in an interactive session with the system, with two sentences involving the verb "to miss", in contexts where it is translated by the Portuguese MWE "sentir a falta":

Introduza uma frase em Inglês ou um comando :
i miss him.

---

DECL1 NP1 PRON1* "i"
VERBI* "miss"
NP2 PRON2* "him"
PUNC1 " ."
---

árvore portuguesa
---

TRAD1 DECL2* NP3 PRON3* "eu"
VERB2* "sinto"
NP4 DET1 ADJI* "a"
DET2 ADJI* "sua"
NOUN1* "falta"
PUNC2 " ."
---

Geração

===> eu sinto a sua falta.
Evaluation of the system

The goal of this section is to describe both our concern for an evaluation procedure and the objective results achieved to date.

It is clear to the MT (and, more generally, to the NLP) community that the evaluation of programs dealing with natural language (and so with an "open" universe) is an extremely difficult task. (See [5]). Not only the demarcation of what is the required output is sometimes not even possible, but the evaluation measures are most of the times subjective in that they vary from user to user.

Specifically concerning MT, numbers like "sentences correctly translated" or "target language words correctly chosen" are actually too poor a measure. The complexity of a sentence can vary considerably, and it is a common danger that a system gets unintentionally tuned for the texts used in its development. (See [4]). Therefore, most of the times the numbers achieved for a particular text are not at all those the system would produce facing a completely new input.

Also, an “incorrect” translation may stem from an awkward realization of the correct meaning by the generation component, a missing term in the bilingual dictionary, or the undue application of structural rules meant for other cases. There should be, in our opinion, performance measures for each component of the system, applicable, furthermore, to each linguistic problem being dealt with.

This would allow not only a much better evaluation of the system, but also, and this is an important issue too, it would help to focalize the development in certain directions: as soon as a certain grade had been achieved for a particular problem, other cases should be solved prior to insisting on the improvement of the first.

We have followed this approach dealing with two particular problems, both considered to be difficult for the pair of languages at hand:

- In structural transfer, the translation of infinitive clauses,
- In lexical transfer, the translation of the verb “to be” (involving the distinction “ser/estar” of some Romance languages).

After considering the problems solved, we ran the system against the sentences of our corpus that contained instances of the situation to test (141 and 166 occurrences, respectively) and measured the following aspects:

- As a black box, which percentage of English examples had gone successfully through the whole translator.
• Focusing then on transfer only, and, having access to the output of the parser, considering it to be correct in all cases, we counted the number of correct translations (about the particular problem at hand, as was already emphasized). The incorrect ones were classified, for future processing, into two categories

- problems not considered by our rules
- actual incorrect result produced by those same rules.

In both cases, the percentage of incorrectness due to each factor was less than 4.5% (3.47 - 2.08 and 1.20 - 4.21 for the infinitives and for the verb "be", respectively). That was considered enough for changing the current focus to other problems, responsible for an overall poorer performance.

Conclusions

This text tried to present the features that seem to us more important in the machine translation prototype we developed. Though in an early stage of its development, we think we can state that MENTOR88/P incorporates many characteristics that provide a considerable power to a machine translation system.

At the same time, we consider that the prototype we have nowadays provides a flexible environment for research and consequent test in this area of automated translation.

References


Sample translations

These are just some examples, but we hope to produce some good ones.

--- estes são apenas alguns exemplos, mas nós esperamos produzir alguns bons.

We are happy to do this demonstration, as we are happy with our work.

--- nós estamos felizes por fazer esta demonstração, assim como nós estamos contentes com o nosso trabalho.

The meeting was in Paris but she wasn't there.

--- a reunião foi em Paris mas ela não esteve lá.

They are happy with him because he is good.

--- eles estão felizes com ele porque ele é bom.
eles estão contentes com ele porque ele é bom.

Functions are assumed to return one value only.
supõe se que as funções devolvem um valor só.
Young people generally like modern music.
os jovens gostam geralmente da música moderna.

For him to be forgotten, they wanted us to erase his name, but we forgot to burn his books.

para ele ser esquecido, eles quiseram que nós apagassemos o seu nome, mas nós esquecemos nos de queimar os seus livros.

I miss the program that is missing from my library.

eu sinto a falta do programa que falta na minha biblioteca.

He was missed by everyone.

foi sentida a sua falta por todos.

But who missed his friend?

mas quem é que sentiu a falta do seu amigo?