Lexical gaps and idioms in machine translation

Diana Santos

IBM-INESC Scientific Group

R.Alves Redol, 9, 1000 Lisboa, Portugal

(internet: dms@inesc.inesc.pt)

Abstract

This paper describes the treatment of lexical gaps, collocation information and idioms in the English to Portuguese machine translation system PORTUGA.

The perspective is strictly bilingual, in the sense that all problems referenced above are considered to belong to the transfer phase, and not, as in other systems, to analysis or generation.

The solution presented invokes a parser for the target language (Portuguese) that analyses, producing the corresponding graph structure, the multiword expression selected as the result of lexical transfer.

This process seems to bring considerable advantage in what readability and ease of bilingual dictionary development is concerned, and to furnish maximal flexibility together with minimal storage requirements. Finally, it also provides complete independence between dictionary and grammar formalisms.

Organization

The general architecture of the MT system is at first described very briefly, emphasizing the features relevant to the full understanding of the problem at hand. Then the problem is presented, and a literature survey given. The solution put forward is then described. Finally, we furnish a detailed example, together with some evaluation results.

The general MT system

The structure of the transfer MT system PORTUGA is illustrated in Figure 1.

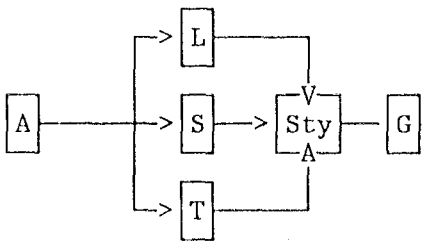

Figure 1. General structure of PORTUGA: A - analysis, G - generation. Transfer: L - lexical, S - structural, T - tense, Sty - style.

The main characteristics of this English to Portuguese translator are:

- the separation between *possible* translation (which may be multiple), and *best* or *chosen* translation (decided in the "style transfer" module).
- Complete independence between English and Portuguese processing. English analysis is performed by PEG[8].
- Bilingual dictionary being kept to a minimum, only the selection conditions for lexical transfer and contrastive knowledge are stored. It should also be mentioned that all information in this dictionary is associated with the translations, and not to the English index, as is usual for lexical transfer in MT.

The reader is suggested to consult [13] or [14] for more details.

The problem

Vagueness, together with non overlapping of semantic fields across different languages is widely known to give rise to lexical gaps, and lexical ambiguity.

For this reason, **lexical transfer**, the process of choosing the correct equivalent(s) for one lexical entry in another language, is one of the most difficult problems that MT has to cope with.

This paper focuses on one aspect of lexical transfer, namely the **possibility to specify complex translations in the target language**. Under this broad description, the use of complex (henceforth **multiword**) expressions, change of part-of-speech required by translation, and collocation restrictions, are meant..

The process of actually choosing which entry (or entries) in the bilingual dictionary is more appropriate, and which information is taken into account for that process has been described elsewhere[13].

Some examples of instances of complex translation (in opposition to simple translation, in the sense of word to one word translation, same part-of-speech, independently of the number of possible choices available) will illustrate the problem in the context of English-to-Portuguese translation.

1-to-N words	
miss	- sentir a falta
miss	- deixar escapar
drop	- deixar cair
kick	- dar um pontapé
tonight	- hoje à noite
graduate	- tirar o curso
N-to-1 words	
have fun	- divertir-se
get up early	- madrugar
fall in love	- apaixonar-se
take advantage	- aproveitar
television set	- televisor
swimming pool	- piscina
N-to-M words	
kick the bucket	- bater as botas
lose one's temper	- perder a paciência

Figure 2. Translation gaps: One-to-many, many-to-one and many-to-many words translation.

Other approaches

In this section I mention related work and alternative solutions that have been proposed and which I find representative of the present day state-of-art. Therefore, primitive approaches such as, for instance, treatment of complex expressions as simple strings will not be surveyed.

Machine translation:

It is acknowledged by outstanding machine translation researchers that there are MT problems which are bilingual in nature. Regarding the problem of lexical transfer, Tsujii[17] states

"we cannot enumerate, by monolingual thinking, different concepts denoted by the verb 'produce'. (...) Only when we are asked to translate sentences into another language, can we try to find appropriate target language words. (...) The above discussion implies that certain 'understanding processes' are target language dependant, and cannot be fully specified in a monolingual manner."

Specifically the problem of translating a word for an expression, is one of the reasons presented by Schenk[15] to use the concept of "Complex Basic Expression" in the description of one language. (A CBE is a basic expression from a semantic point of view, i.e., it corresponds to a basic meaning, and a complex expression from a syntactical point of view.)

"Expressions that are not idiomatic, but that consist of more than one word can be handled by means of a complex basic expression in order to retain the isomorphy."

This approach is related to the theoretical requirement of the Rosetta MT system to build isomorphic grammars for the several languages dealt with by the system.

This implies that, in this framework, it is the set of all languages in presence that defines what is a basic meaning. Like this, it is possible to dispense altogether with structural transfer,

"Structural transfer is not necessary, since idioms are based onto basic meanings."

In general, however, most MT systems do not make the analysis phase dependant on the target language(s), and therefore it is usual to see statements like the following[4]:

"the lexical rule must mark words with the corresponding parts of speech,", in the *"cases where a source language entry must be rephrased in the target language as an ad-hoc combination of words which does not form a lexical entity"*.

It is unclear, however, how much complexity of the result can be handled, or how much syntactical transformation the new expression can suffer.

Some hybrid approaches can also be found in [9], this time apparently putting the burden on the generation phase:

"The second substep of German morphological generation is the application of German word list transformations. (...) these rules can also be used to handle non-compositional translations. For example, "for example" can translate compositionally into ***für Beispiel****, and then a GPHRASE rule can convert this to* ***zum Beispiel****."*

However, the more comprehensive way to deal with this problem, without changing analysis accordingly, is the one exemplified by Isabelle[7]. It was developed

"a special language called LEXTRA, which makes it easier to state the type of tree transformations required by lexical transfer. (...) LEXTRA takes as data an explicit description of the admissible tree structures, and guarantees that any tree it receives or creates is indeed an admissible tree".

Similar solutions can be found in the Japanese-to-English system of [10], which handles both lexical gaps and changes of part of speech:

"One can specify not only the English main verbs but also arbitrary phrases governed by the verbs as constants", allowing for variables and complex patterns in the lexical rules. *"one can provide lexical rules directly in GRADE, and attach them to specific items. (...) One can specify specific tree transformations in GRADE"*.

and in the old ITS translation system[12]:

"The contents of these records" (dictionary entries) *"include transfer language statements which performs the necessary transfers as well as other referential information."*

Parsing

In parsing, idioms have to be considered. A recent paper on idiom processing[1] lists some of their relevant properties (my rephrasing):

- *usual existence of ambiguity between literal and idiomatic readings,*
- *frequent discontinuity of idioms,*
- *applicability of regular application of syntactic rules (like adverb(s) or auxiliary verb(s) insertion),*
- *applicability of 'transformations' to the idioms proper, (like passivation or relativization).*

The difference between "non-literal reading" and "idiomatic" reading of an expression is also pointed out. Metaphoric readings are proposed to be parsed by usual rules. The advantages of submitting idioms to "regular" syntactic rules and even to 'transformations', whenever possible, are emphasized.

A more extreme view can be found in Gazdar et al[3], who ignore idioms as far as syntax is concerned:

"no additional devices need be added to the syntax in accounting for the peculiarities of fixed expressions". Since not only idioms can be assigned internal syntactic structure, but an internal semantic structure as well, as *"all syntactically active idiomatic expressions have a metaphorical basis"*.

However, radically different views can be found for instance in [5], where, in the lexicon-grammar approach, the concern with the representation of compound words (adverbs, verbs, nouns) makes Gross establish a classification according to their syntactical shape, ranging from several degrees of variation, from completely frozen ("at night") to having parts completely free ("organize in one's honor").

This author suggests that finite automata be attached to a given entry in order to describe the compound variation.

"The variations of form we have enumerated can be partly handled by attaching a finite automaton to a given entry, and this automaton will describe the main grammatical changes allowed."

In between, the need to store several pieces of information concerning idioms is acknowledged by Stock[16], such as

undergoing passivization, weight in the whole idiom, remover of the idiom interpretation, semantic value, etc.

This system stores idioms as "further information concerning words", divided in two cases, "canned phrases" and "flexible idioms", the latter being stored under the 'thread' of the idiom.

Based on the claim that *"the flexibility of an idiom depends on how recognizable its metaphorical origin is"*, one of the goals is to *"integrate idioms in our lexical data merely as further information concerning words (as in traditional dictionaries)"*.

Generation

Finally, research in natural language generation has also contributed to clarify and furnish solutions to the problem.

Clearly, generation is one of the issues in a machine translation system. However, work in generation per se usually presupposes the existence of an unambiguous 'concept' representation, and so the problems begin with the correct stating of an idea in one particular language. In this framework, it is clear that one key concept is that of "collocations", or how lexical items combine in a particular language.

In these systems, it is advocated (see for instance [6]) that in the specialized 'semantic' dictionary *"storing the possible lexicalizations of a 'concept' in a given language (...) the possibility of combining lexemes in collocations"* should also be stored, specifically in the entries for the bases (which determine the possible collocates: a collocation is a pair base-collocate).

A remark of utmost importance can be found in [11], during the description of the DIOGENES generation system:

"collocational relations are defined on lexical units, not meaning representations".

Summing up

The literature survey above supports some of our assumptions, namely that

- there are problems which are bilingual in nature, and cannot therefore be properly dealt with in only one language;
- there is not a clear distinction between what should be accounted for as an idiom, a metaphorical use of a word or a collocation. The boundaries between collocational restrictions, metaphorical readings and idioms are blurred and may even not be pertinent to the automatic treatment of language.
- to translate correctly, it is often necessary to use expressions instead of single words. Those expressions can moreover give origin to complex structure changes, possibly discontinuous.

Our approach

We are interested in solving the problem of translating one expression into another expression, no matter whether the need arises because of a lexical gap, a collocation difference or an idiom not literally translatable.

Therefore, we treat all these three problems the same way, namely, considering them as instances of a contrastive lexical transfer problem in the scope of machine translation.

We must emphasize that we are only interested in expression-to-expression translations when the literal ones are not acceptable. This stems from the fact that there is a considerable number of fixed expressions which do not require any special processing, as can be seen in the following list, with examples taken from several languages:

(E) parents and children
pais e filhos
(E) ladies and gentlemen
senhoras e senhores
(F) monter la moutarde au nez de
subir a mostarda ao nariz de
(F) attendre un enfant
esperar uma criança um filho
(E) take into account
tomar em conta
(I) prendere il toro per le corna
pegar o touro pelos cornos
(E) in good hands
em boas mãos

Figure 3. Literally translatable idioms: E-English, F-French, I-Italian

Our solution

Given that the target expressions can be arbitrarily complex, we impose no restrictions whatsoever on their form or structure, and give unlimited power to the device intended to cope with them.

On the other hand, it didn't appeal to us to have to store, for each pair source-target entries, the full structural transformation implied, as in the most powerful approaches mentioned above (cf. [10] and [7]). This approach gives origin to very heavy dictionaries, with a lot of redundancy, moreover, since there may be similar transformations repeated to many entries. On the other hand, not only the dictionary becomes very difficult to understand and modify, (requiring someone who knows the "programming" language used), but also it makes it tightly coupled to the structural representation and/or particular linguistic formalism and options used in the machine translation system, in both analysis and generation.

We chose thus a different method that

- allows for maximal readability
- is independent of the linguistic (and programming) decisions of the whole MT system (being only concerned with lexical transfer)
- provides as much power as any unrestricted (tree or graph) transformation language

The method proposed consists then of using the target string as the result value in the bilingual dictionary, therefore keeping it independent of whatever structure it should be assigned, and invoking a target language parser that builds the structure required, on the fly.

Another advantage of the process above is that the new structure is dynamically built only when it is necessary (that is, when it corresponds to the chosen translation).

On the other hand, no separate (and redundant) lexical rules need be written in the dictionary, as the very same grammar is used for all multiword target expressions. The grammar should be a "twin" of that used in the analysis

phase, that is, it should obey the same formalism and linguistic options in order for them to be compatible.

A detailed example

For the sake of clarity, a full example will be presented, regarding the word *miss*, in its meaning of *to feel sorry or unhappy at the absence or loss of (someone or something)* (Longman). The Figure 4 shows an abridged form of the entry for *miss* in the bilingual dictionary. The information for choosing among the several possible translations was omitted and will not be discussed here. The examples presented will be in any case those that correctly trigger the translation *sentir a falta* (literally, "feel the lack").

```
miss(VERB CHTPOSS (EVP sentir a falta))
miss(VERB CHTPOSS (EVP ter saudades))
perder(VERB)
faltar(VERB)
miss(VERB (EVP deixar escapar))
menina(NOUN)
```

Figure 4. Dictionary entry for "miss": EVP stores the Portuguese string to be used as translation.

The first thing that should be exemplified, is that, after the choice of the multiword translation, the Portuguese grammar is invoked, building a equivalent graph fragment to the translation of "feel the lack". This graph fragment is then conveniently inserted in place of the one for *miss*.

```
i miss you.
-----------------------------------------------------
DECL1 NP1    PRON1* "i"
      VERB1* "miss"
      NP2    PRON2* "you"
      PUNC1  "."
-----------------------------------------------------
árvore portuguesa
-----------------------------------------------------
DECL2 NP3    PRON3* "eu"
      VERB2* "sinto"
      NP4    DET1   ADJ1*  "a"
             DET2   ADJ2*  "tua"
             NOUN1* "falta"
      PUNC2  "."
-----------------------------------------------------
Geração
 = = = = > eu sinto a tua falta .
```

Figure 5. A simple example.

With this simple example, it can be seen that some structural manipulation took place (converting the English direct object pronoun into a Portuguese possessive adjective - triggered by the CHTPOSS marker in Figure 4), and that the words taking part in the multiword expression were conveniently inflected (in this case, only the verb).

More complex processing can clearly take place, as is exemplified in Figure 6.

```
I 'll always miss people i like .
 = = = = > eu sentirei sempre a falta de pessoas de
 quem eu gosto .
I miss the man who was here .
 = = = = > eu sinto a falta do homem que esteve aqui .
he was missed , but who missed her ?
 = = = = > foi sentida a falta dele , mas quem é que
 sentiu a falta dela ?
he was the one who most missed his father .
 = = = = > ele foi o que sentiu mais a falta do seu
 pai .
they were the ones who were least missed .
 = = = = > eles foram os de quem se sentiu menos a
 falta .
I miss having you in the neighborhood .
 = = = = > sinto a falta de te ter na vizinhança .
I forgot missing you .
 = = = = > esqueci-me de sentir a tua falta .
I forgot to miss you .
 = = = = > esqueci-me de sentir a tua falta .
```

Figure 6. Several examples of "miss" translated as "sentir a falta": Complex tenses, passive, relative clauses, distinction between third person singular and others, adverb position, etc.

As *ter saudades* is also a valid translation for *miss* in the same context as *sentir a falta*, this choice belongs to style transfer. Follows the output of the system in that case:

```
i miss that time.
 = = = = > eu tenho saudades daquele tempo.
```

Figure 7. Another style alternative: "miss" translated by "ter saudades".

Other problems

It remains to be shown how the other problems mentioned above are solved in this framework. We begin by change of part-of-speech, and continue by identifying source language (English) multiword expressions, which then comprehend the remaining cases, namely collocations and equivalence of distinct idioms.

Change of part-of-speech

The change of part-of-speech should be transparent as far as the dictionary is concerned, being the assignment of the correct interpretation performed by the Portuguese parser.

```
thank(VERB (NPOS AJP) (PREPO OBJECT a)
      (EVP obrigada))
agradecer(VERB (PREPO for por))
agradecimento(NOUN (PREPO for por))
obrigado(NOUN)
```

Figure 8. Abbreviated entry for the word "thank": Since the string "Obrigada" has three possible interpretations according to the Portuguese parser, NPOS stores the phrase type to select.

Only when there are more than one parse for the target expression and the one to choose implies a change of part-of-speech needs this to be stored in the bilingual dictionary, as can be seen in Figure 8 above.

Note: Mori Rimon pointed out to us that in cases of highly ambiguous target languages, as is the case of written

Hebrew, the indication of which syntactical alternative, when different from the source one, could be needed very frequently, therefore reducing the economy we are asserting. We can only answer that while for English-to-Portuguese translation that structural marking is very rarely used, further testing with different language pairs must be done in order to assess or deny the universality of this method. Namely languages whose translation would require an extensive part-of-speech change should be tested.

Follows a very simple example of the case discussed above:

```
thank you.
-----------------------------------------------------
DECL1  VERB1*  "thank"
       NP1     PRON1*  "you"
       PUNC1   "."
-----------------------------------------------------
árvore portuguesa
-----------------------------------------------------
DECL2  ADJ1*   "obrigada"
       PP1     PREP1   "a"
               PRON2*  "ti"
       PUNC2   "."
-----------------------------------------------------
Geração
 = = = = > obrigada a ti .
```

Figure 9. Change of part-of-speech.

This example would be improved if the whole phrase "thank you" were translated by "obrigada", but here we want to show the simplest case.

It should also be mentioned in passing that whenever there is a generalized part-of-speech change on syntactical grounds, that is not done through lexical transfer, but in the structural transfer, as is the case of adjectival English present participle clauses.

MWE-to-MWE translation

Considering the general problem of identifying source language multi word expressions, the philosophy we propose is similar. (We are indebted to Stephen Richardson for this suggestion.) The implementation is however not yet done, so what will be described in the rest of this chapter is only a proposal.

We consider that source expressions should be identified as a bilingual requirement too, and therefore this process should take place (only if needed) during transfer. If the identification succeeds, the whole phrase would then be replaced by the corresponding Portuguese translation, be it a word or a complex expression.

The next examples illustrate how the bilingual dictionary would look like:

```
thunder(NOUN (MWE thunder and lightning)
             (EVP relâmpagos e trovões))
```

Figure 10. Collocation differences: The same device used for many-to-many translation can be used when, for instance, the order must be reversed.

```
kick(VERB (MWE kick the bucket)
          (EVP bater as botas))
```

Figure 11. Example of a many-to-many words translation: The MWE feature corresponds both to the context required in order to choose that particular translation, and to the piece of English to replace.

Some numbers

In order to evaluate the interest and need for taking this problem into account in machine translation, the following measures were performed, regarding an English-to-Portuguese MT dictionary roughly containing 500 English entries and 2400 Portuguese translations. Only the case one English word to several Portuguese words was taken into account.

```
No. of English entries with EVPs : 80.
Total number of EVPs : 152.

No. of verbs translated by EVPs : 60.
No. of nouns translated by EVPs : 47.
No. of adjectives translated by EVPs : 14.
No. of adverbs translated by EVPs : 19.

No. of entries whose first translation is an EVP : 27.
No. of nodes whose correct translation is an EVP : 42.
```

Figure 12. Some relevant numbers

In order to guarantee impartiality of the numbers presented, the criteria for selecting the English entries, and the actual translations, bore no relationship whatsoever with the problem mentioned in this paper.

The numbers arrived at, however impressive they may be, should nevertheless not be confused with percentages of occurrence in actual text. On the contrary, there is some relationship between a rarely used word in one language and a set of words to express it in another language.

However, we still consider that the numbers above unequivocally demonstrate that this problem cannot be ignored in any real machine translation system.

As for the actual testing of the proposed method, we ran the system on two test corpora, the first, regarding the verb "miss", including several different syntactic environments (see Figure 6), and the second containing several different instances of 1-to-N translations:

I stood in the doorway .
= = = = > estive de pé na soleira da porta .
I dropped the camera while packing .
= = = = = > deixei cair a máquina fotográfica enquanto estava a fazer as malas .
I missed the sunset tonight .
= = = = = > senti a falta do pôr do sol hoje à noite .
The filmstar kicked her agent .
= = = = = > a estrela de cinema deu um pontapé ao seu agente .
Watch the dog !
= = = = = > toma cuidado com o cachorro !
I bicycled and did not do my homework .
= = = = = > andei de bicicleta e não fiz o meu trabalho de casa .
A then officer would not borrow a uniform .
= = = = = > um oficial do tempo não pediria emprestado um uniforme .
Did I trouble you when I yellowed your shirt ?
= = = = = > causei-te transtorno quando tingi de amarelo a tua camisa ?

Figure 13. Several examples of 1-to-N translation.

Even though no thorough broad-coverage translation tests have been performed, we believe these results can assess not only the feasibility but also the flexibility of the method proposed.

Conclusion

The approach presented in this paper handles in the same way the problems of lexical gaps, collocation requirements in different languages, and non-literal translation of idioms.

Considering them a bilingual problem, the transfer phase was assigned as the proper place for them to be treated.

The method presented has as advantage minimal storage required and the least computation (only on demand) of the several structures involved. Also, it only makes use of one single comprehending parser for the target language, instead of developing particular solutions to particular problems.

The way the dictionary was conceived brings with it considerable readability, making it independent of the linguistic and programming formalisms used in the other modules of the translation system. Its format can, moreover, make it very easy to inherit information from human-readable bilingual dictionaries. Hand-coding by an expert is not required.

References

[1] Abbeillé, Anne and Yves Schabes. 1989 "Parsing Idioms in Lexicalized TAGs", *Proceedings of the Fourth European Conference of the European Chapter of the Association for Computational Linguistics*, 10-12 April 1989, Manchester, UK.

[2] Beaven, John L. and Pete Whitelock. 1988 "Machine translation using isomorphic UCGs", *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 22-27 August, 1988.

[3] Gazdar, Gerald, Ewan Klein, Geoffrey Pullum and Ivan Sag. 1985 *Generalized Phrase Structure Grammar*, Basil Blackwell.

[4] Golan, Igal, Shalom Lappin and Mori Rimon. 1988 "An Active Bilingual Lexicon for Machine Translation", *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 22-27 August, 1988.

[5] Gross, Maurice. 1986 "Lexicon-Grammar: The Representation of Compound Words", *Proceedings of the 11th International Conference on Computational Linguistics*, Bonn 1986, pps 1-6.

[6] Heid, Ulrich and Sybille Raab. 1989 "Collocations in Multilingual Generation", *Proceedings of the Fourth European Conference of the European Chapter of the Association for Computational Linguistics*, 10-12 April 1989, Manchester, UK.

[7] Isabelle, Pierre. 1984 "Machine Translation at the TAUM Group", *Machine Translation Today: The State of the Art*, Margaret King, ed., 1987.

[8] Jensen, Karen. 1986 "PEG 1986: A Broad-coverage Computational Syntax of English", IBM Research Report RC draft, Feb 1986, T.J.Watson Research Center, Yorktown Heights, NY 10598.

[9] McCord, Michael C. 1989 "Design of LMT: A Prolog-Based Machine Translation System", *Computational Linguistics*, Vol. 15, No. 1.

[10] Nagao, Makoto and Jun-ichi Tsujii. 1986 "The transfer phase of the Mu Machine Translation System", in *Proceedings of COLING'86*, ACL, pps 97-103.

[11] Niremburg, Sergei and Irene Niremburg. 1988 "A Framework for Lexical Selection in Natural Language Generation", *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, 22-27 August, 1988.

[12] Richardson, Stephen D. 1980 "A High-Level Transfer Language for the BYU-TSI Interactive Translation System", M.A. Thesis, Brigham Young University.

[13] Santos, Diana. 1988 "A fase de transferência de um sistema de tradução automática do inglês para o português", Tese de Mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa.

[14] Santos, Diana. 1988 "An MT prototype from English to Portuguese", *Proceedings of the IBM Conference on Natural Language Processing*, October 24-26, 1988, Thornwood, pps 122-133.

[15] Schenk, André. 1986 "Idioms in the Rosetta Machine Translation System", *Proceedings of the 11th International Conference on Computational Linguistics*, Bonn 1986, pps 319-324.

[16] Stock, Oliviero. 1989 "Parsing with Flexibility, Dynamic Strategies, and Idioms in Mind", *Computational Linguistics*, Vol. 15, No. 1.

[17] Tsujii, Jun-Ichi. 1986 "Future directions of machine translation", *Proceedings of the 11th International Conference on Computational Linguistics*, Bonn 1986, pps 655-668.

Acknowledgements

This paper greatly benefited from the comments of Jan Engh, Stephen Richardson and Mori Rimon, and from Paula Newman's critical reading of an earlier version.

I am therefore grateful to them and to all members of the IBM-INESC Scientific Group for their support and discussion.