An Open Transfer Translation

Jorge Kinoshita

Escola Politécnica da Universidade de São Paulo
Departamento de Computação e Sistemas Digitais (PCS) - Brazil
email: jkinoshi@pcs.usp.br URL: http://jk.pcs.usp.br

Abstract. We are developing an English-Portuguese Transfer Machine. The transfer machine operates in three phases: the analysis phase is done according to a dependency grammar, the transfer phase is done according to a transfer dictionary and the generation phase conjugates the Portuguese words. The user interface is done through the web. Our system is "open" because the user can view intermediate structures generated by the system and change the database system in order to correct the text during the revision process.

1 Introduction

We are developing an English-Portuguese [1] transfer machine based on some presuppositions: (1) we do not need to get perfect texts translations. Our aim is to speed the translation process performed by humans. (2) two related words (a bigram) in an English sentence defines a good context and hence, it is possible to associate a good translation to the bigram. A database of bigrams and their translations (which we call, "a transfer dictionary") can be used in the analysis phase (the more bigrams are found in the analysis, the better) and in the transfer phase (in order to translate bigrams). (3) instead of revising the text that was automatically translated, the user should revise the transfer dictionary. When the user changes one entrance in the dictionary, the system shows how the entire text is modified due to this alteration. For instance: the system can wrongly translate "home page" to "página (page) do(of) lar(home)" The correct translation is "home page" (The Brazilian Portuguese keeps the English term). If the translation "home page" $\rightarrow$ "home page" is added to the dictionary, all the occurrences of "home page" will be automatically revised; then speeding the translation process.
The translation process that is performed automatically consists of :
(1) link the words of an English sentence in pairs (by some dependency grammar). It is possible to associate a relation-tag (kind of linkage) to a bigram. For instance in "spread-wing", the relation-tag is VB/NN that means "verb-object".
(2) translate the bigrams according to the dictionary.
The translation process that is performed manually consists of: (1) select an wrong translation of an word or bigram and correct it. (2) submit the correction to the system which will correct the necessary portions of the text. (3) repeat the two former steps until no improvement can be made. (4) correct manually

all sentences of the text (this will not change the database system). For instance: the anaphora, ellipses and idiomatic expressions are difficult to be handled by bigrams.

2 The Transfer Dictionary

In the transfer dictionary, it is possible to declare the translation of:
(1) a **bigram** (two words). Example: "give up → desistir". In this example two words are translated as a single word. It is possible to add and delete words and even change words position.
(2) a **partial bigram** (one word and its relation-tag to the head). Examples:
(A) "like -VB/NN → gostar-de" : the verb "like" when linked to its object is translated as "gostar" plus the preposition "de" that is inserted before the translated object. If no object exists, the preposition "de" is not inserted.
(B) "it -PRP\R → -" : the pronoun "it" behaving as a subject is deleted in Portuguese translations.
(3) a **single word** according to its tag. Example: "like VB → gostar".
It is also possible to rearrange the sentence according to the Portuguese syntax based only on the relation-tag. For instance, in Portuguese, adjectives generally follow the noun and this is handled looking at the relation-tag JJ\NN (adjective-noun).
The dictionary follows a syntax that is beyond the scope of this article.

3 The Translation

Translation is done automatically in 3 phases:
- **analysis**: link each word to its head (one word, the root of the sentence, does not have a head). For instance, the sentence "the boy sold a new house" is analyzed as in Table 1.

Table 1. Analysis Result

pos	head	tag	word
1	2	DT	the
2	3	NN	boy
3	0	VBD	sell sold
4	6	DT	a
5	6	JJ	new
6	3	NN	house

Table 2. Transfer Result

pos'	pos	head	tag'	word'
1	1	2	DT	o
2	2	3	NN	menino
3	3	0	VBD3s	vender
4	4	6	DTf	um
5	6	3	NNf	casa
6	5	6	JJf	novo

where:
pos = word position in the sentence; *head* = head of the word; *word'* = Portuguese word in infinitive form.
The structure shows how words are linked to each other. For instance, "house"

is the object of “sold” because the word 6 has the word 3 as its head. The tags follow the TREEBANK [2] notation.

- **transfer**: The input is the structure generated by the analysis and the output is a sequence of words in Portuguese but in the radical form as it is shown in Table 2.

The translation of bigrams is done in a top-down direction because it is easier to handle the movement of words and phrases around the head-word. The statements used from the transfer dictionary are presented in Table 3.

Table 3. Rules

$JJ \backslash NN$: % + %
a DT : *um*
boy NN : *menino*
house NN : *casa.f*
new JJ : *novo*
sell VB : *vender*
the DT : *o*

- **generation**: After the transfer phase, the words must be conjugated according to their tags and the result is: “o menino vendeu uma casa nova”.

Revision. The revision is done interactively with the system: **1-system** displays all the translated text. **1-user** selects the sentence of the text that he intends to fix. **2-system** displays the analysis and transfer structures. **2-user** selects any word of the transfer structure (for instance, the word “house”). A selection of one word implies in the selection of a bigram: the word and its head (ex: house and sell). **3-system** displays all possible entrances of the transfer dictionary related to the selection; in this example, to the word “house”, to the bigram “sell-house”, to “house” behaving as an object and to “sell” behaving as a transitive verb. The system also displays blanks to be filled by the user if there is lack of some entrance in the dictionary (which normally happens, specially for the bigram translation). For instance, let us suppose that the translation of “house” were not declared. Its translation appears as a blank enabling the user to provide new information. **3-user** corrects the system database. **4-system** looks in the text for all possible cases where it can be applied. This process does not imply in redoing all translation again.
The system keeps all intermediate structures of the parser and transfer phases in a database. This is done because: 1) as the system is designed to be in internet, it must attend many users. A user must not block any process during his interaction to the intermediate structures of his text. 2) any user alteration that affects the intermediate structures can be checked very fast; otherwise the system would have to generate all the structures again. 3) the intermediate structures

in a database can be used as a noted corpus providing information for future programs.

4 Conclusion

Our transfer mechanism is based on simple operations on bigrams. The simplicity of the formalism suggests that it could be applied (or easily adapted) to projects that have an analysis similar to that provided by dependency grammars (ex: Link Grammar [3]). The user needs to be prepared to use the system. He needs to know the rule syntax for deleting English words, adding Portuguese words, reordering the words in Portuguese, etc. In a future version, a more user-friendly interface will be created. Today, the system allows many users changing the database and working on the same translation. The problem is that a user can damage the translation of other users. In a future version, we are planning to have some mechanism of controlling versions in order to recover to previous states. Today, the system registers in a file the operations performed by all users. This file will be used to recover to previous states.
System builder and contact: the author.
System category: research vehicle.
System characteristics: translation of technical texts (given in HTML or plain text). The system is treating 50 sentences per minute in its automatic translation (Pentium MMX 200MHz: local test). This result must be seen with care: the output still requires a lot of revision work and there were many small phrases (titles) that were counted as sentences.
Resources: word-net1.6 [4], and GDBM files of around 1.7 Mbytes for dictionaries and a small grammar of 50Kbytes.
Hardware and software: Pentium PC; operating system: Linux RedHat 4.2.
Functionality description: Today our system does not speed the translation process. The revision phase requires a lot of work because: (1) our database is still very poor. Today the purpose is to create mechanisms to control it. (2) the parser needs to be improved.

acknowledgement: FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo

References

1. KINOSHITA, J.: Aspectos de Implementação de Uma Ferramenta de Auxílio à Tradução Inglês-Português. Tese de Doutorado, Departmento de Computação e Sistemas Digitais da Escola Politécnica da Univesidade de São Paulo, Brazil (1997)
2. PENN TOOLS. http://www.cis.upenn.edu/ adwait/penntools.html, (1998)
3. SLEATOR, D. D. K.; TEMPERLEY, D. : Parsing English with a Link Grammar. Pittsburgh, Canegie Mellon University, sleator@cs.cmu.edu (1991)
4. MILLER, G.A. ET AL.: Introduction to WordNet: an on-line lexical database. Princeton University, http://www.cogsci.princeton.edu/ wn/, (1993)