A Bigram Based Transfer Dictionary

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 present a transfer dictionary model that explains how words and pair of words can be translated from one language to another in a concise way. The dictionary is being used in a English-Portuguese translation machine project, however, the transfer dictionary model is generic enough to be applied to other projects as well.

1 Introduction

The machine translation based on the transfer approach operates in three steps: 1) **analysis** : maps sentence As into a syntax tree Ast. 2) **transfer** : maps the structure Ast into another syntax tree Bst, through the use of some transfer grammar. 3) **generation** : maps the syntax tree Bst into the sentence Bs. The transfer grammar is a set of rules heavily dependent on the kind of structure resulted from the analysis. For instance [1]:

$[gov : know, subj : \$1, obj : \$2[NP]] => [gov : conhecer, subj : \$1, obj : \$2]$.

This rule specifies that if the verb *"know"* has a subject and an object of type NP (noun phrase) then it must be translated as *"conhecer"* in Portuguese (*connaitre* in French) instead of *"saber"*. This rule, in particular, can not be applied to the sub-sentence : *"... in order to know that man"* because the right side of the rule does not match the syntax tree due to the lack of subject (*"know"* is in infinitive). This rule should focus only on two words (a bigram): the verb *"know"* and its object (ex: *"man"*). We assume that most of the transfer rules can (and should) be written focusing only on bigrams because they define a good context and tend to be translated in the same way. In this paper, we present a simple notation to deal only with bigram translation. These rules can be organized as a dictionary.

We are building an English to Portuguese transfer dictionary for a machine translation project. The translation process is:

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 each bigram. For instance in *"know-man"*, the relation-tag is +VB_NN that means *verb-object*.

2) translate the bigrams according to the dictionary and assemble them, building the translated sentence S.

3) put the words of S in their normal form based on their radical form.

2 The Transfer Dictionary

The translation of bigrams deals with many aspects as it is shown in the examples: (a) *yellow - man : homem - amarelo*; (b)*yellow - woman : mulher - amarela*; (c) *give - up : desistir*; (d) *like - orange : gostar de - laranja*; (e) *when - translating : ao traduzir*. A bigram is generally translated as another bigram (ex: a, b, e). The word *"yellow"* is translated to *"amarelo"* or *"amarela"* depending on the genre of its head. Another peculiarity is that the adjective appears before the noun in English and in the opposite order in Portuguese. In example (c), *"give-up"* is translated as one single word. In example (d), the Portuguese verb *"gostar"* is not directly linked to its object and requires the preposition *"de"*. In example (e), the verb *"translating"* was modified from gerund to infinitive in Portuguese.
We propose a set of rules to deal with these facts in bigram translation. These rules can be organized in a dictionary. There are 3 kinds of entrance in the dictionary: 1) *a bigram* (ex: *give - up*). 2) *a single word and a relation-tag* (ex: *know - +VB_NN*) and 3) *a single word and its tag* (ex: *man - NN*). First, we are going to explain the rule syntax in order to translate two words and then, show how this same syntax can be generalized to translate one word according to its relation-tag.

2.1 Translation of Bigrams

The two words of a bigram are in a dependency relation. For instance: in the pair (concern, with) we can say that *"with"* is subordinated to *"concern"*. We can associate a relation-tag to each bigram. The relation-tag is derived from the two morphological tags of each word. For instance, the relation-tag of (concern, with) is +VB_PREP (verb and preposition). These two tags (ex: VB and PREP) are ordered according to the sequence of words in English sentences. The sign "+" before the first tag shows that the first word is the head, otherwise the sign "-" is used. For instance, the relation tag of (big, house) is -JJ_NN. The relation-tag does not follow exactly the TREEBANK notation [4]. For instance: -NN_VBP, -NN_VBD, etc. are mapped to -NN_ROOT in order to give the notion of *"subject"*.
The translation rule syntax of a bigram is:

$H\ H^T\ S\ RT : H'\ S'$

For instance:

$concern\ VB\ with\ +VB\_PREP : preocupar\ com$

where : H = head-word; H^T = tag of H according to the TREEBANK notation; S = subordinated-word; RT = relation-tag; H' = translation of head-word; S' = translation of subordinated-word.
The pattern ($H\ H^T\ S\ RT : H'\ S'$) is incremented in order to specify the following operations:
a)**Order declaration**. When a bigram is translated, the order of the 2 words can be different in Portuguese. For instance: *"Ethernet protocol"* is translated as *"protocolo Ethernet"*. The order can be specified by adding the sign "+" to S'

(otherwise, the English order is maintained): S'+ means that H' comes after S' and +S' means that H' comes before S'. For instance:

protocol NN *Ethernet* $-NN\_NN$: *protocolo* $+$ *Ethernet*

In the transfer phase, the order declaration imposes a movement of not only the dependent-word but also of the entire phrase bellow it. For instance: *"a very fast car"* is translated as *"um carro muito rápido"* (a car very fast). The fact that S' *"rápido"* (fast) was moved to a position after H' *"carro"* (car) made the entire phrase *"muito rápido"* (very fast) be moved after *"carro"*.

b)**suppression of S' or H'**. There are cases that S' or H' is an empty string. This is declared by the "minus" sign in the place of the translated word (generally the S'). For instance: The case of translating *"there + to be"* as *"existir"* is easily handled by the rule:

be VB *there* $-EX\_VB$: *existir* $-$

c)**insertion of word**. A word in Portuguese can be inserted in the translation. For instance,*"like"* is generally translated as *"gostar"* plus the insertion of the preposition *"de"* . For instance, *"like chocolate"* is declared as:

like VB VB_NN- *chocolate* : *gostar* $<-$*de chocolate*

The notation $H' <-P$ means that P must be placed before the phrase that has S' as its head.

2.2 Translation of One Word Based on the Relation-Tag

Instead of writing many bigrams (know, English), (know, John), etc. and their translations, it is possible to omit the object of *"know"*:

know VB $<VB\_NN-$: *conhecer*

The general statement is:

$H\ H^T < RT : H'$ or $H\ H^T < RT : H'\ S'$ (head in focus) and

$S\ S^T > RT : S'$ or $S\ S^T > RT : H'\ S'$ (subordinated-word in focus).

For instance, in Portuguese it is common to suppress the translation of *"it"* when behaving as the subject (-NN_ROOT) of some sub-sentence. This fact is expressed as:

it $PRP > -NN\_ROOT : -$

The *"no translation"* **Symbol - % :** It is possible to refer to some translated word W' (H' or S') without imposing any translation. This is helpful when we want to specify some attribute of W' such as its position in the translated sentence or its morphological tag. When the "%" symbol replaces W' in a statement S, then S does not impose any translation to W. Other rules will take care of translating W. For instance: the rule ($H\ H^T < RT : H'\ S'$):

boy $NN < -JJ\_NN$: *menino* $+\%$

means: any word referred by "%" (JJ/adjective) must come after *"menino"*. It can be applied to $(yellow, boy) => (menino, amarelo)$.

Changing Tags : By default, the Portuguese tag follows the English tag. Sometimes, it is necessary to set a new tag to S' or H' using the "::" symbol. Ex:
$when\ WRB < WRB\_VBG- : ao\ \% :: VB$
Ex: (when, writing) => (ao, escrever) (*writing* /VGB: gerund, *escrever* /VB: infinitive).

2.3 Single Word Translation

Single words are also declared in the dictionary:
The syntax is:
$W\ W^T : W'$
Example:
$bidirectionality\ NN : bidirecionalidade : f$
The set of tags in Portuguese is different from English. For instance, in Portuguese, nouns are classified according to its genre. By default, the genre is masculine. In this case, ":f" means "feminine". The determiner and adjective must agree in genre and number with its head. The agreement is made automatically during the transfer phase.

2.4 Syntax Rearrangement Observing Only the Relation-Tag

Other transfer rules that does not depend on words can be created guided by the above pattern. The rule syntax is:
$S^T << RT : H'\ S'$
or:
$H^T >> RT : H'\ S'$
In these rules, H' and S' are generally expressed by % . For instance:
$NNP >> -NN\_NN : \%\ +\%$
Ex: *"TCP/IP protocol"* becomes *"protocolo TCP/IP"*.

3 Using the Dictionary in the Transfer Phase

In our machine translation project the translation is done in 3 steps:

1) **Analysis**. Each word is linked to its head (one word, the root of the sentence, does not have a head). For instance, *"the boy sold a new house"* is analyzed according to Table 1.
The first column designates the word position in the sentence and the second column its head. For instance, *"house"* is the object of *"sold"*.

2) **Transfer**. The graph of words as knots and dependency relations as arches is generally a tree. The transfer phase takes all the bigrams from the root to the

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

leaves and translates each bigram according to the transfer dictionary. First it tries to translate the bigram if it is declared in the dictionary; otherwise it tries to translate H or S based on the relation-tag RT of the bigram and, then, finally it tries to make a syntax rearrangement based only on RT. The next step is to translate all words that did not fit any rule for bigrams. An example of the transfer phase output is in Table 2.

Table 2. Transfer result

pos	pointer	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

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 were:
$NN << -JJ\_NN : \% + \%$
$aDT : um$
$boyNN : menino$
$houseNN : casa : f$
$newJJ : novo$
$sellVB : vender$
$theDT : o$

3) **Generation**. The words must be conjugated according to their tags and the result is: *"o menino vendeu uma casa nova"*.

4 Conclusion

Operations such as exchanging word places, suppressing/inserting words are very common in the transfer phase. In spite of it, there are few works in the literature about this theme. For instance, [6] is an interesting work about the English to Japanese translation and presents a transfer grammar that is more complex than the dictionary presented here. The main contribution of this paper is to propose a division between general transfer rules that operates on the syntax tree and transfer rules that operates only on bigrams. Simpler rules for bigrams were possible because, even though two words in one language can be translated as a different number of words in another language, our notation enables a bigram to be always mapped into another bigram. This approach can not cover all transfer rules. For instance, the trigram *"as well as"* requires other rules. This formalism is being applied to our machine translation project with success. The dictionary enables a simple transfer system, for instance, it is not necessary any complex grammar for sentence generation. The simplicity of the formalism suggests that it could be applied (or easily adapted) to projects that has an analysis similar to that provided by dependency grammars (ex: Link Grammar [5]). The project is in evolution and we intend to make the transfer dictionary work with the WordNet [3]. It is expected that it is possible to have just one translation for each meaning in WordNet.
We believe that the transfer dictionary notation works fine due to the great tendency in human languages to group related words together as it was observed in many theories ([7], [8], [2]).

acknowledgment : partially supported by FAPESP.

References

1. DOUG, A.: Transfer. working paper, Department of Language and Linguistics, University of Essex, UK, (1997) 1–8
2. MELCUK, I. A. : Dependency syntax: theory and practice. New York, State University of New York Press, 1988. apud SLEATOR (1991)
3. MILLER, G.A. ET AL.: Introduction to WordNet: an on-line lexical database. Princeton University, http://www.cogsci.princeton.edu/ wn/, (1993)
4. PENN TOOLS. http://www.cis.upenn.edu/ adwait/penntools.html, (1998)
5. SLEATOR, D. D. K.; TEMPERLEY, D. : Parsing English with a Link Grammar. Pittsburgh, Canegie Mellon University, sleator@cs.cmu.edu (1991)
6. TSUTSUMI,T.: Wide-range restructuring of intermediate representations in machine translation. Computational Linguistics, v. 16, n. 2, (1990) 71–78.
7. RADFORD, A. : Transformational grammar: A first course. Cambridge, Cambridge University Press, (1989).
8. RAPOSO, E.P. : Teoria da gramática: A faculdade da linguagem. Lisboa, Editorial Caminho S.A., (1992).