SIMPÓSIO 46

Propostas Integração de Metodologias para o Estudo do Processo Tradutório

COORDENAÇÃO:

Professora Ana Luísa Leal
FSH/DP – Universidade de Macau, Macau
analeal@umac.mo

Professor Fábio Alves
FALE – Universidade Federal de Minas Gerais, Brasil
fábio-alves@ufmg.br

Professor Wong Fai
FST/DCIS – Universidade de Macau, Macau
derekfw@umac.mo

HUMAN ANALYSIS OF MACHINE TRANSLATED DOCUMENTS BETWEEN PORTUGUESE CHINESE MACHINE AIDED TRANSLATION SYSTEM AND ONLINE TRANSLATION SYSTEMS

Fai WONG277, Francisco OLIVEIRA277, Sam CHAO[1]

ABSTRACT: There are many machine translation tools available in the market nowadays for handling the translation between different languages. Among them, Google Translate is widely used in the society. Although they are easy to use and quick for obtaining the translation, they may not generate good results when dealing with the localization of documents in the city. As a result, this paper presents an evaluation of localized documents translated by the Portuguese-Chinese Machine Aided Translation System (PCT) and Google Translate in terms of different criteria based on the linguistics' point of view, including the fluency and adequacy. Sentences are carefully selected for this purpose, and these contain Macau specific words, phrases extracted from local newspapers and government documents. In order to further understand typical problems in these translation systems, a summary of the orthographic, lexical, and syntactic problems is explained in details. The collection of the information is done by an online PCT evaluation interface.

KEYWORDS: Machine Translation, Localization, Manual Evaluation

1. Introduction

The automatic translation of languages from one to another through the use of computers is becoming more and more attractive, not only from the researchers, but also to people who need to have translations in their daily work.

In Macau, a city with Asian and Western cultures, it always considered as a multiple language society, with three writing and four spoken languages. Currently, the official languages are Portuguese and Chinese. Nowadays, there are various machine translation systems in the market being used by many enterprises and educational entities. The core design of these systems is divided into three paradigms. Example-based MT (EBMT) (Brown, 1996; McTait, 2001) and Statistic-based MT (SMT) (Brown et al., 1990; Lopez, 2008) paradigms rely on corpora which comprise of bilingual texts. EBMT analyzes different pieces of bilingual examples in the extraction and combination of phrases for generating the

1 Faculty of Science and Technology, University of Macau
Av. Padre Tomás Pereira, Taipa, Macau
{olifran, derekfw, lidiasc}@umac.mo

translation. On the other hand, SMT rely on probabilities estimated between the translation of words and the ordering of sentences extracted from the corpora. On both approaches, the accuracy of the translation is highly dependent with the size and information of the corpora. Rule-based MT (Bennett et al., 1985) approach is based on a set of linguistic grammar rules for handling the translation.

Among different available translation systems, Google Translate is widely used nowadays. It relies on the SMT paradigm for handling the translation task between several languages based on huge multilingual corpora. However, since it uses probabilities estimated in handling the translation task, for some words/phrases, and specific terms which are specific to the city, the system may not always generate a correct result. Another system, which is used in the Macau society for handling the translation task between Portuguese and Chinese, is called PCT System (Oliveira et al., 2010). It is based a hybrid translation system which is based on multiple translation engines. Moreover, it applies Translation Corresponding Tree (TCT) (Wong et al., 2006) for searching and matching the fragments between bilingual texts, and Constraint Synchronous Grammar (CSG) (Wong et al., 2006) for establishing the relationship between the languages simultaneously based on semantic constraints.

This paper provides a deeper study on some typical problems which will occur in the translation results generated by MT systems with focus to domain related to the Macau society. These problems are classified as orthographic errors, lexical translation problems, and syntactic issues. On the other hand, the translation quality is accessed in terms of fluency and adequacy (White et al., 1994) with the help of an evaluation interface developed by NLP²CT research group.

The rest of the paper is organized as follows. Section 2 introduces the problems found during the study of translated sentences generated by these systems. Section 3 introduces the online evaluation interface developed in ranking the quality of the translated results. Preliminary evaluations conducted in terms of translation quality are detailed in Section 4, followed by conclusion and future improvements.

2. Problems studied in automatic translation results

This paper focuses on different aspects in terms of the sentences translated, including incorrect lexical selection, translation problems related to numbers, special characters translation issues, and syntactic ordering problems. In the next sections, we highlighted some examples identified during our study period. However, there are many other cases in which the translation results generated are not qualified.

2.1 Lexical Selection Problems

When the domain is related with Macau documents, we noticed that for some terminologies, which are widely used in Macau, are not correctly translated by Google. Some of the found examples are shown in Table 1.

The terminology errors occurred in Google are mainly due to the inherent properties of SMT systems. Since probabilities of the whole sentence are considered in the selection of the most suitable translation equivalent, phrases being translated according to different context may generate different results. Another possible reason is due to the corpora which are used to train the system. As the information in the trained corpora cannot cover all the words and phrases for all domains, it may not be always possible to get the correct translation.

On the other hand, for PCT systems, since they have a specific lexicon for Macau terminologies, translations related with local streets, companies, entities, etc., the translation quality can be guaranteed, and always produce the same translation across the text.

Table 1. Local phrases translation generated by Google and PCT System

Macau Related Terminologies	Source Sentence	Translated Result	
		Google	PCT System
Street name	Avenida do Ouvidor Arriaga	阿裡亞加大道監察員	雅廉訪大馬路
Names of entities, government departments, companies	Banco Nacional Ultramarino	北京師範大學	大西洋銀行
	posto fronteiriço do COTAI	在邊防哨所的路氹	路氹邊檢大樓
	澳大	Austrália	Universidade de Macau
Monuments	Portas do Cerco	門	關閘

2.2 Numbering and Special Characters Translation Problems

In this study, we found that the translation of some numberings and the understanding of special characters meaning cannot be handled correctly. Some of the examples are revealed in Table 2. In the first example, the number "71/2011" is translated as a fraction by Google Translate since it treats the symbol "/" wrongly under this context. On the other hand, it seems that the system cannot recognize the meaning of "º" for the second example. Comparing with the PCT System, since it is targeted for translating administrative documents, often used symbols are carefully handled with their correct translation.

It might happen for some special characters, PCT system cannot recognize them correctly, as shown in the last example. It can be easily solved by adding new entries for these symbols in the knowledge base.

Table 2. Numbering translations generated by Google and PCT System

Source Sentence	Translated Result	
	Google	PCT System
n.º 71/2011	第二千〇一十一分之七十一	第71 / 2011
1.º trimestre	1季度	第一季
«Universidade de Macau»	“澳門大學”	《澳門大學.

2.3 Syntactic Ordering Problems

Syntactic Ordering problems often occur in MT systems when the knowledge applied is out of domain. According to different context, different types of errors may occur. In this paper, we summarized two common issues studied from the analyzed results. The first issue is related to the wrong groupings related to punctuations when long sentences are being translated. An example is shown in Table 3. The system grouped the phrase "portátil, câmara de vídeo" instead of using the comma delimiter as a boundary to handle the translation task. Correct groupings may avoid this type of error, and the reason for this problem is probably due to the probabilities estimated during the training phase.

Table 3. Example of Punctuation problem in the translation

Source sentence	... , écran gigante e portátil , câmara de vídeo ...
Translated Result (Google)	... ， 屏幕巨大 ， 便攜式攝像機 ...
Reference Translation (English)	... , big and portable screen , video camera ...

The second issue is related to wrong groupings related to noun phrases in shorter sentences. In the example shown in Table 4, besides the unknown phrase "Portas do Cerco" generated by Google Translate, some syntactic orderings are not correct. The translation of "Península de Macau" is translated as "半島澳門" without properly swapping the words

SIMPÓSIO 46

“半島” and “澳門”. We noticed that for longer sentences, the handling for Google Translate may have different translation results since it heavily relies on the probabilities estimations between words/phrases in the context. On the other hand, for PCT systems, although similar phrases translation may appear disregarding the sentence length, since the rules defined cannot cover all the sentences, specially for long ones, it also suffers from the syntactic ordering problem.

Table 4. Example of Noun phrases translation problem

Source Sentence	Translated Result	
	Google	PCT System
Além das Portas do Cerco - a fronteira norte da Península de Macau	除了門 -北部邊界的半島澳門	除了關閘 - 澳門半島北的邊界

Evaluation Interface

The study conducted in this paper is based on an online evaluation interface developed by the NLP²CT research group.

Figure 1. Screenshot of the translation interface

The whole process is divided into three phases. Once users are logged in to the system, in the first phase, they will input sentences to the system and wait for the translation results, as shown in Figure 1. A segmentation function is provided to check either correct segmentations are made at the sentence level.

In the second stage, users will evaluate the quality of the generated results, as shown in Figure 2, which are classified as: *Readable*, *Need Modification*, and *Unreadable*. *Readable* refers to translations with good fluency and adequacy. *Need Modification* cases are referring to sentences which require some modifications in order to make them understandable. *Unreadable* results refer to cases which are not understandable, inadequate, or not fluent.

Figure 2. Screenshot of the evaluation interface

At last, if any problems are found, evaluators may describe them in the remarks area. The objective in to collect common errors found in these results so that they can be improved by adding proper knowledge into the MT systems.

Experiments and Discussions

Different experiments are performed to evaluate the quality of online MT system. The test set consists of Macau local documents from different domains, including Administrative Bulletin documents, Macau News, University of Macau News, and tourism related documents. The quality of the translation results is evaluated by human assessments, classified as Readable, Need Modification, and Unreadable.

In the first experiment, we randomly select 1100 sentences from these test sets, and they are evaluated by three linguistics. In the average results shown in Table 5, 84% of the translations require modifications, and only 15.7% of them are readable.

Table 5. Human Assessments results of MT generated translations

	Readable	Need Modification	Unreadable
Assessments	15.7%	84%	0.3%

In the second experiment, in the sentences which require modifications, 20% of them are further studied. In order to have a more fair judgment, cross validations are considered. In other words, after the test set is evaluated by one linguistic, the same set will be re-evaluated by another. The number of errors committed based on the criteria discussed previously are counted, and the results are shown in Table 6.

Table 6. Errors committed in translations which require modification

Type of Error	Percentage
Grammar Related	10%
Unknown Words/Phrases; No Translation	46%
Wrong Lexical Selection	36.7%
Others (Numbering, Special Characters, etc)	7.3%

In the results, we found that most of the errors are related to the lexicon (82.7%), either translation of words are wrongly selected, specific phrases translations are unknown, or there are no translation generated. Since many of the documents used in this evaluation are mainly based on local news and administrative related fields, online translation system generated a number of errors related to lexicon compared with the other types of errors.

In the last experiment, a comparison is made between the translation quality of Google Translate and PCT system. In the news domain, we found that Google generates better results than PCT system, especially the readability even it is not adequate. On the other hand, for administrative documents, since PCT system is initially targeted to help professional translators in the government, the translation quality of PCT system is better than the Google Translate. No doubt, there are also many cases in which both systems perform the same, either having well translated results or totally wrong translations.

5. Conclusion

This paper presents a preliminary study on the translations generated by Google Translate and PCT System with focus to the Macau documents domain. Different criteria are accessed, including lexical selection problems, numbering and special characters handling issues, and syntactic ordering problems. On the other hand, the translation quality is

evaluated by linguistics which are classified as Readable, Need Modification, and Unreadable. We noticed that Google Translate may not always guarantee Macau terminologies compared to PCT System. From the experiments, we found that Macau Bulletin documents are better handled by PCT Systems. For longer sentences translation, we found that Google Translate has a better readability compared to the PCT System even though their adequacy are the nearly the same.

In the future, we planned to continue on this study in the identification of more typical errors generated by these systems from several domains related to Macau. Moreover, these conclusions can serve as valuable clues for improving the MT systems.

Acknowlegements

This work was partially supported by the Research Committee of University of Macau under grant UL019B/09-Y3/EEE/LYP01/FST, and also supported by Science and Technology Development Fund of Macau under grant 057/2009/A2.

References

Bennett, W.S., Slocum, J.. 1985. *The LRC Machine Translation System*. Computational Linguistics 11(2-3), 111-121.

Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.. 1990. *A Statistical Approach to Machine Translation*. Computational Linguistics 16(2), 79-85.

Fai Wong, Ming Chui Dong, and Dong Cheng Hu. 2006. *Machine Translation Based on Translation Corresponding Tree Structure*. Tsinghua Science and Technology, 11(1), pp. 25-31.

Fai Wong, Ming Chui Dong, and Dong Cheng Hu. 2006. *Machine Translation Using Constraint-Based Synchronous Grammar*. Tsinghua Science and Technology, 11(3), pp. 295-306.

Francisco Oliveira, Fai Wong, Sam Chao, Chi-Wai Tang. 2010. *Portuguese Chinese Machine Aided Translation System*. The Anthology of Selected Papers from FIT 6th Asian Translator's Forum, Macau.

Google Translate. Available at: <http://translate.google.com>. Accessed in August, 2011.

Lopez, A.. 2008. *Statistical Machine Translation*. ACM Computing Surveys, Vol. 40, No.3, Article 8.

McTait, K.. 2001. *Translation Pattern Extraction and Recombination for Example-Based Machine Translation*. PhD Thesis, Centre for Computational Linguistics, Department of Language Engineering, UMIST.

White, J., O'Connell, T., O'Mara, F.. 1994. *The ARPA MT evaluation methodologies: evolution, lessons, and future approaches*. Proceedings of the 1994 Conference, Association for Machine Translation in the Americas, Columbia, Maryland.

ENSEMBLE LEARNING ON PORTUGUESE POS TAGGING

Xiao-Dong ZENG[278]
Sam CHAO[2]
Fai WONG[278]

ABSTRACT: Ensemble learning, also known as multiple classifier system, can combine the predictions from multiple base classifiers/learners altogether to conclude a final decision. It has been proven that ensemble learning is a simple, useful and effective meta-classification methodology. SBCB (Selecting Base Classifiers on Bagging) is a selective based ensemble learning algorithm which is able to select an optimal set of classifiers amongst all base classifiers in determining the final result. SBCB equips with an optimization process that is capable of selecting a suitable number of optimal classifiers among all base classifiers automatically, in which, diversity and accuracy are considered as two major selection criteria. In this paper, we built a part-of-speech (POS) tagger for Portuguese based on SBCB learning algorithm as a case study to further investigate the effectiveness and performance of SBCB algorithm. The problem of POS tagging is a practical issue in natural language processing (NLP), especially in the development of a machine translation system. The performance of the POS tagging may interference the subsequent analytical tasks in the translation process, and thereafter affects the translation quality. The POS tagging task can be regarded as a classification problem. Features such as the surrounding context of ambiguous candidates, *n*-gram information, lexical items and linguistic clues are used and automatically extracted from the Portuguese annotated corpus. The empirical results reveal the effectiveness of SBCB algorithm on POS tagging.

KEYWORDS: Ensemble learning, SBCB, POS Tagging, Classifier selection

1. Introduction

In machine learning and pattern recognition, classification refers to an algorithmic procedure for assigning a given piece of input data into one of a given number of categories (Huang et al., 2006). That process generally depends on a so call classifier trained by algorithm itself. General classification algorithms, such as Decision Tree, Neural Network, Bayesian and so on, only create a single classifier in training phase. Recent years, a novel classification algorithm category named Ensemble Learning that can generate and combine multiple classifiers has been proposed. Ensemble learning, is also called multiple classifier system, employs multiple learners and combines their predictions to a classification task. By combining multiple base classifiers, it is aiming to obtain a more accurate classification decision at the expense of increased complexity.

SBCB (Selecting Base Classifiers on Bagging) algorithm proposed in our previous work (Zeng et al., 2010) is a selective ensemble learning method. SBCB algorithm utilizes an optimization process to select the optimal classifiers (subset of classifiers) from the original candidates to be the final base classifiers. It means that not all the trained classifiers would be used as the members of targeted classifier in decision phase. In the design of the optimization process, both the accuracy and diversity of the possible classifier candidates are taken into consideration as the selection criteria. In (Zeng et al., 2010), the elementary evaluation of the algorithm based on multiple existing benchmark datasets from UCI machine learning repository were fully conducted and the evaluation results reveal that SBCB learning algorithm does outperform the conventional generic bagging method in terms of learning accuracy and complexity. Based on such advantages of SBCB approach, in this paper, it was applied to the construction of part-of-speech tagger for Portuguese. Part-of-speech tagging, in natural language processing, is the process of automatically and correctly assigning part-of-speech or lexical class marker to each word in a sentence. It is a vital but hard task in reality, since large number of words' POS in a text tends to depend on both the definition and context of the word, which generates so many

2 Faculty of Science and Technology, University of Macau
Av. Padre Tomás Pereira, Taipa, Macau
nlp2ct.samuel@gmail.com, {lidiasc, derekfw }@umac.mo

SIMPÓSIO 46

hard-handling tagging ambiguities. In this paper, we tried to resolve the ambiguity by using classifier trained on SBCB learning algorithm. Since assigning part-of-speech to a word can be regarded as classification problem, where the label or class is POS and the instance is the word plus its neighboring context.

The structure of this paper is as follows: section two overviews the related work of ensemble learning and classical POS tagging algorithms. In section three, the idea of SBCB learning algorithm is reviewed. Section four describes the process of constructing POS tagger for Portuguese based on SBCB learning algorithm. Section five demonstrates the result of experimental works including the evaluation of classifier and the tagger, followed by a conclusion to end this paper.

2. Related Work

2.1 Ensemble Learning

Recent years, more and more researchers are concerning on ensemble learning. The reason is that ensemble learning tends to yield better performance and generalization ability than that of individual classification algorithm. (Dietterich, 2002) Ensemble learning, is also named as multiple classifier system, employs multiple learners and combines their predictions to a classification task. By combining multiple base classifiers, it is aiming to obtain a more accurate classification decision at the expense of increased complexity. According to (Dietterich, 2002), there are three types of reasons (statistical, computational and representational) why a classifier ensemble might be better than a single classifier.

In general, an ensemble model contains four layers: input layer, feature layer, base classifier layer and fusion layer. Input layer mainly focuses on providing various kinds of datasets. Feature layer creates different feature sets using the output of input layer. Base classifier layer mainly emphasizes how to create base classifiers and design the structure of classifiers. Fusion layer is in charge of combining the prediction results from a variety of base classifiers using some fusion rules, ensemble rules, or combinative rules, etc.

The representative ensemble methods are Adaboost and Bagging. Adaboost (Freund et al., 1995), which sequentially generates a series of base learners, where the training instances that are wrongly predicted by a base learner will play more important role in the training of its subsequent learner. Bagging (Bootstrap Aggregating) (Breiman, 1996) does not sequentially generate base learners, but parallel bootstrapping resamples in diffident datasets to produce diverse base classifiers.

2.2 POS Tagging Algorithm

Most POS tagging algorithms fall into one of three categories: rule-based tagger, stochastic tagger and transformation-based tagger.

Rule-based tagging algorithms generally involve a large database of hand-written disambiguation rules. The earliest algorithms (Haris, 1962), (Klein et al., 1963), (Greene et al., 1971) for this category generally contains two main stages, where the first stage is in charge of assigning each word a list of potential part-of-speech by using a dictionary and while the second stage is to window down this list to a single part-of-speech for each word by using a large number of hand-written disambiguation rules. Obviously, the performance of the tagger totally depends on the size and quality of disambiguation rules. The collection of rules, however, is manually difficult and time-consuming. Although the tagging algorithms, described in (Voutilainen, 1995), use a rich and large dictionary and disambiguation rule set, the performance cannot be remarkable improved.

Stochastic or probabilistic tagging algorithms generally resolve tagging ambiguities by using a training corpus to compute the probability of a given word having a given tag in given context. A popular stochastic tagging algorithm is known as the Hidden Markov model (HMM) tagging algorithm (Merialdo, 1994). In this tagging approach, the task becomes the POS sequence identification problem, which is to take the tag sequence that gives the highest probability.

Transformation-based tagging or Brill tagging introduced by Brill (Brill, 1995) shares features of both before-mentioned tagging algorithms. The tagging processing for the ambiguous word of this tagger also relies on rules which are automatically induced by learning component based on tagged training corpus.

3. SBCB Learning Algorithm

SBCB algorithm is one of our previously work on ensemble learning proposed in (Zeng et al., 2010). SBCB stands for Selecting Base Classifiers on Bagging, which embeds an optimization process into generic Bagging algorithm that can automatically infer an optimal set of classifiers. The whole process of SBCB algorithm is illustrated in Figure 1. Initially, an original dataset and a base classification learning algorithm are declared. During the training phase, it bootstraps resampling from original dataset to create m training sets. It in turns is used to train m classifiers using the training datasets based on a base training algorithm. Hence, classifiers C_1, C_2, …, C_m are created. In conventional Bagging training algorithm, the process will cease and the whole training task is completed. But in SBCB, an optimization process is introduced to the candidates of trained classifiers. After optimization process, portion of classifiers combination C_1, C_2, …, C_n ($n \leq m$) is selected, and the whole training process is completed thereafter. In the classification of an instant, voting rule is applied to combine the predictions generated from the selected classifiers to give a final prediction.

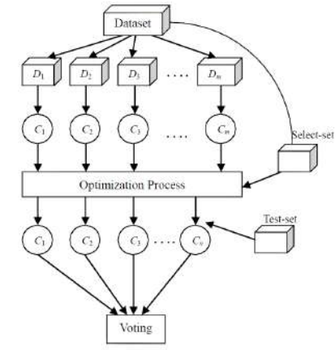


Figure 1. Process of SBCB algorithm

4. POS TAGGER BASED ON SBCB LEARNING ALGORITHM

In tagging (Portuguese) sentences, the objective is to process or handle three kinds of words: unambiguous, ambiguous and unknown words. Unambiguous words refer to the words that have only one part-of-speech. According to the statistics gathered on several Portuguese corpora, there are about 50% unambiguous words. Thus, the process of tagging those unambiguous words can be easily achieved by looking up their POSs from a dictionary. For an ambiguous word, it has more than one possible POS that recorded in the dictionary and its particular POS depends on the context. While an unknown word is not included in the dictionary. In order to determine the tagging information for the last two types of words, our tagger is designed as a classifier by using SBCB learning algorithm that trained on collection of ambiguous and unknown words in terms of feature based representations.

SIMPÓSIO 46

The methodology of constructing a part-of-speech tagger for Portuguese based on SBCB learning algorithm is illustrated on Figure 2. From the figure, we can observe that a completed tagger should consist of a dictionary and a classifier. The dictionary, lexicon and its POS, is used to pre-assign a POS tag to the unambiguous word considering that which is the only POS. In our work, the dictionary is built from tagged corpus incrementally. While the SBCB classifier is created to determine the proper tagging information for the ambiguous and unknown works in the sentence.

In tagging an input sentence, the model will go through the following processing steps. First, for each word of sentence, dictionary will be used for looking up the POS information. For those, which have more than one possible POS will be regarded as ambiguous words, and together with the unknown words, will then be processed by the classifier. In this step, features, including the active word as well as its surrounding context, concerning the words are extracted and represented as a feature-based instance (feature value vector). Then, the instances are fed into the classifier for predicting the target tag information.

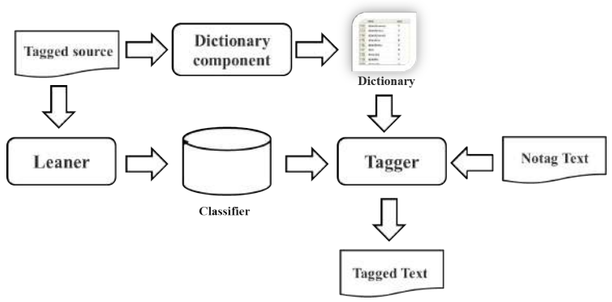


Figure 2. Construction of POS Tagger

4.1. Dictionary construction

The dictionary is the basic component of the entire POS tagger. The objective of the dictionary is to facilitate checking which word in a sentence is ambiguous word or unknown word and at the same time assigning the POS to unambiguous word. As shown on the Table I, there are three columns of the dictionary: index id, word and POS. Every single entry in dictionary is recording word and its pre-occurred POS. The entries in dictionary are collected from tagged corpus automatically and incrementally.

Table I. Entries of dictionary

Id	Word	POS
1	abas	N
2	abastada	V
...	...	...

4.2. Classifier induction

As before-mentioned, our POS tagger is built based on dictionary and classifier. This part will introduce how to induce or train the classifier based on SBCB learning algorithm. Like other classification algorithms, the classifier of our tagger is trained on the feature-based instances extracted from the tagged corpus. Each ambiguous word is represented by a set of features. As shown on the Table II, 17 features are defined to describe an ambiguous word. The features that do best to the classifier are selected through experiments. This includes the part-of-speech information around the word, n-gram information and morphological information (suffix, prefix and etc.). Therefore, the features of an active word are transformed and described with an 18-dimension vector <f_0, f_1, ..., f_{16} | label>, where f_0, f_1, ..., f_{16} represents feature values and label indicates the true POS of word. Figure 3 shows the examples of instances extracted from the tagged sentence.

Table II. Feature collection

No.	*Feature*	*Description*
0	W	Local word
1	W^0	Previously-1st word
2	W^{-1}	Previously-2nd word
3	W^{-2}	Next-1st word
4	P^{-1}	POS of previously-1st word
5	P^{-2}	POS of previously-2nd word
6	P^{1}	POS of next 1st word
7	Suf_3	Suffix (3)
8	SL_3	Start with lower case?
9	SU	Start with upper case?
10	CC	Contain any capital letter?
11	AL	All lower case?
12	AU	All upper case?
13	CN	Contain Number?
14	CP	Contain period?
15	CH	Contain hyphen?
16	CO	Contain other symbol?

Sentence: *Alguma/Q vez/N **se/SE** havia/HV de/P ver/VB **a/D** vaidade/N sem/P lugar/N ./.*

Word: se
Instance: <se, vez, Alguma, havia, N, Q, HV, null, Y, N, N, Y, N, N, N, N, N | SE >
Word: a
Instance: <a, ver, de, vaidade, VB, P, N, null, Y, N, N, Y, N, N, N, N, N | D>

Figure 3. Example of feature value extraction

5. EXPERIMENTS

In this section, two experiments are designed to evaluate our proposed model. In the first experiment, we compared the performance of SBCB algorithm and other classification algorithms on pre-collected ambiguous word dataset. The second experiment focuses on the evaluation of the whole part-of-speech tagger.

Tycho Brahe corpus, which is a parsed corpus of historical Portuguese, will be utilized to conduct the experiments (Charlotte et al., 1999). It contains 52 texts (more than 40,000 sentences) from the Institute of Mathematics and Statistics of the University of São Paulo. The sentences have been manually tagged with POS and syntactic features at the University of Campinas in the lines of the Penn-Helsinki Parsed Corpus of Middle English (Helena et al., 1999). In the Tycbo Brahe corpus, there are using 154 different level tags. These tags were projected on a smaller system of 41 tags by maintaining the first level of tags.

5.1. Evaluation of classifier

As mentioned before, the main responsibility of classifier is the disambiguation for word's POS, which is assigning a proper POS to ambiguous words. In order to investigate how effectiveness of SBCB algorithm on tagging the ambiguous words, an empirical comparison for SBCB and other classical classification algorithms were performed in this section.

SIMPÓSIO 46

To setup the experiment, we first prepare a dataset of ambiguous words in which every single word is converted into a multi-feature instant as described in section 4.2. According to the requirement, we have extracted 88,273 instances from 31,318 sentences in Tycho Brahe corpus.

The algorithms that involved in this experiment consist of two single classifiers: C4.5 (Decision Tree), Naïve Bayes; and three ensemble classifiers: Adaboost, Bagging, and SBCB. For the evaluation strategy, 10-fold cross-validation was used to carry out the evaluation for different classifiers. In the 10-fold cross-validation, dataset is divided into ten datasets. One of them is used as the testing dataset to evaluate the performance of the classifier that was trained on the rest of the other nine datasets. The process iterates though the all combinations (ten times), and the experimental results are averaged.

Table III. Accuracy of different algorithm on dataset (%)

Classifier	C4.5	Naïve Bayes	Adaboost	Bagging	**SBCB**
Accuracy	71.2	70.8	83.2	83.0	**84.5**

Table III describes the experiment results of five classification algorithms on the collected ambiguous word dataset. From the results, it is observed that the three ensemble learning methods give a better performance in comparing with that of single classifiers, in which SBCB algorithm outperforms the other algorithms. According to the above experimental result, we believe that SBCB algorithm is quite suitable for use as a classification algorithm for the task of part-of-speech tagging.

5.2. Evaluation of tagger

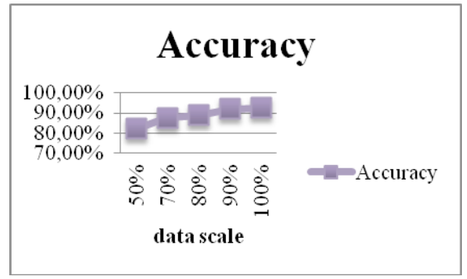
In this section, we will focus on the evaluation of the entire part-of-speech tagger. The evaluation contains two phases: training phase and testing phase. Training phase is to construct the tagger, collecting the content of dictionary and training a classifier. In the testing phase, the trained tagger is use to tag the unseen sentences from the test set.
Table IV. Evaluation dataset of Tycho corpus

Type	***Sentences***	***Tokens***
Train dataset	31,318	638,437
Test dataset	4,095	84,029

Again, the evaluation experiment is conducted using Tycho Brahe corpus in which 31,318 sentences (638,437 tokens) were used for training and the rest of 4,095 sentences (84,029 tokens) were prepared for the testing as illustrated on Table IV. In order to better understand the effectiveness and generalization ability of our tagging approach, we capture the results of taggers trained on different setting of the training data. The corresponding results are given on Table V. The result shows that as the increase of the training data, the accuracy of tagger is getting higher. Since larger training dataset not only can enrich the dictionary, but also can provide more instances to train up a stronger classifier.

Table V. Evaluation of tagger trained on different scale data

Scale	***Correct tagged***	***Accuracy***
50%	68,895	**82.0%**
70%	73,609	**87.6%**
80%	74,702	**88.9%**
90%	77,340	**92.0%**
100%	77,609	**92.4%**



6. Conclusion

In this paper, we are seeking to further evaluate the proposed SBCB algorithm through a practical application of part-of-speech tagger for Portuguese sentences. For this purpose, we firstly evaluated the SBCB learning algorithm whether it is an adaptable classification algorithm for part-of-speech tagging, through the empirical comparison between SBCB and other classification algorithms. From the experimental result, we found that SBCB learning outperforms other classification algorithms. Then, we proposed to use the SBCB algorithm to build a Portuguese part-of-speech tagger. The preliminary result shows that our tagger can achieve 92.4% accuracy on the Tycho Brahe corpus.

Acknowlegements

The authors are grateful to Research Committee of University of Macau for the funding supports of our research, which under the reference RG060/09-10S/11R/CS/FST.

References

Bernard Merialdo. 1994. *Tagging English Text with a Probabilistic Model*. Computational Linguistics, pp.155-171.

Breiman Leo. 1996. *Bagging predictors*. Machine Learning. Vol 24, No. 2, pp.123–140.

Charlotte Galves and Helena Britto. 1999. *A Construção do Corpus Anotado do Português Histórico Tycho Brahe: o sistema de anotação morfológica*. IV PROPOR, Evora: University of Evora, pp. 55-67.

Dietterich Thomas. *Ensemble learning*. 2002. The Handbook of Brain Theory and Neural Networks, Second Edition.

Eric Brill. 1995. *Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging*. Computational Linguistics, pp. 543-566.

Freund Yoav, Robert E. Schapire. 1995. *A decision -theoretic generalization of on-line learning and an application to boosting*. Proceedings of the 2nd European Conference on Computational Learning Theory, pp. 23-37.

Greene Bob and Gerald Rubin. 1971. *Automatic grammatical tagging of English*. Department of Linguistics, Brown University, Providence, Rhode Island.

Helena Britto, Charlotte Galves, Ilza Ribeiro, Marina Augusto, and Ana Paula Scher. 1999. *Morphological annotation system for automatic tagging of electronic textual corpora: from English to Romance languages*. 6th International Symposium of Social Communication, Santiago de Cuba, pp. 582-589 Huang Te-Ming, Vojislav Kecman, Ivica Kopriva. 2006. *Kernel Based Algorithms for Mining Huge Data Sets*. Supervised, Semi-supervised, and Unsupervised Learning, Springer-Verlag, Berlin, Heidelberg, pp. 260

Klein Sheldon and Robert Simmons. 1963. *A computational approach to grammatical coding of English words*. Journal of the Assiciation for Computing Machinery, 10(3), pp. 334-347.

Voutilainen. 1995. *Morphological disambiguation*. Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text, pp. 165-284.

Zellig Sabbettai Harris. 1962. *String Analysis of Sentence Structure*. Mouton, The Hague.

Zeng Xiao-Dong, Sam Chao, and Fai Wong. 2010. *Optimization of Bagging Classifiers Based on SBCB Algorithm*. International Conference on Machine Learning and Cybernetics, Qingdao, China, pp. 262-267.

A UTILIZAÇÃO DE INFORMAÇÃO LINGUÍSTICA EM ABORDAGENS BASEADAS EM APRENDIZAGEM AUTOMÁTICA PARA O PROCESSO DE TRADUÇÃO

Paulo Miguel Torres Duarte QUARESMA[3]

RESUMO: O processo de tradução automática de textos tem sido objecto de abordagens bastante distintas: desde propostas que visam a análise profunda dos textos e a representação da informação num formato independente da Língua, até abordagens baseadas na análise superficial aos textos e às palavras que os compõem, passando por abordagens estatísticas do processo de tradução. Neste artigo analisa-se uma abordagem mista e integradora ao processo de tradução automática: por um lado, efectua-se uma análise linguística aos textos, efectuando a sua etiquetagem morfo-sintáctica, a identificação de entidades mencionadas, uma análise sintáctica parcial e uma análise semântica parcial; por outro lado, recorre-se a técnicas de aprendizagem automática, baseada em métodos estatísticos, para criar modelos de tradução entre pares de Línguas, utilizando corpora paralelos disponíveis nestas Línguas. A integração entre estas duas metodologias é efectuada através da utilização dos resultados da análise linguística como input de métodos de aprendizagem supervisionada. Esta estratégia de integração de informação linguística com técnicas de aprendizagem automática revelou bons resultados quando aplicada ao problema de classificação automática de textos em Língua Portuguesa e pretende-se, neste artigo, analisar a sua possível aplicação à temática da tradução automática de textos em Língua Portuguesa.

PALAVRAS-CHAVE: Tradução Automática; Aprendizagem automática; Linguística computacional

Introdução

A tradução automática de textos é uma área de investigação que tem sido objecto de análise ao longo das últimas décadas. As abordagens propostas têm sido bastante distintas e têm sofrido variações significativas ao longo do tempo, acompanhando a evolução e os desenvolvimentos das áreas de Linguística Computacional e de Inteligência Artificial: desde propostas que visam a análise profunda dos textos e a representação da informação que veículam num formato independente da Língua, até abordagens baseadas numa análise superficial aos textos e às palavras que os compõem, passando por abordagens puramente estatísticas do processo de tradução.

Neste trabalho pretende-se analisar e discutir a utilização de uma abordagem mista e integradora ao processo de tradução automático: por um lado, efectua-se uma análise linguística aos textos, efectuando a sua etiquetagem morfo-sintáctica (POS-tagging), a identificação de entidades mencionadas (nomes, instituições, locais, tempo), uma análise sintáctica parcial (identificando sintagmas nominais e verbais e os seus principais constituintes) e uma análise semântica parcial (recorrendo a ontologias externas e à teoria de representação do discurso DRT); por outro lado, recorre-se a técnicas de aprendizagem automática, baseadas em métodos estatísticos, para criar modelos de tradução entre pares de Línguas, utilizando corpora disponíveis nestas Línguas. A integração entre estas duas metodologias é efectuada através da utilização dos resultados da análise linguística como input do processo de aprendizagem automática. Esta estratégia de integração revelou bons resultados quando aplicada ao problema de classificação automática de textos em Língua Portuguesa (Gonçalves 2009) e é passível de ser aplicada à temática da tradução automática em geral e de textos em Língua Portuguesa, em particular.

O artigo encontra-se estruturado da seguinte forma: na próxima secção é descrita a arquitectura proposta; na secção seguinte são apresentadas as ferramentas de extracção de informação linguística utilizadas no âmbito do processamento de textos em Língua Portuguesa; posteriormente, é feita uma descrição breve das várias técnicas de aprendizagem automática que podem ser utilizadas no processo de tradução; finalmente, são discutidos alguns dos problemas existentes e possíveis linhas de trabalho e investigação futura.

3 Universidade de Évora, Escola de Ciências e Tecnologia, Departamento de Informática, Rua Romão Ramalho nº 59, 7000 Évora, Portugal, pq@uevora.pt

SIMPÓSIO 46

Arquitectura

De forma a abordar de uma forma integrada o problema da tradução automática de textos, propõe-se o recurso a uma arquitectura modular, em que numa primeira fase os textos são analisados através de ferramentas de processamento de Língua Natural, sendo o resultado dessa análise utilizado por algoritmos de aprendizagem automática supervisionada, com o objectivo de melhorar os modelos construídos e os resultados obtidos.

De uma forma esquemática, a abordagem proposta pode ser descrita pelo esquema da figura 1:

Textos PLN Aprendizagem Automática Tradução Proposta

Os textos, escritos na Língua "origem", são analisados por ferramentas de processamento de Língua Natural, de forma a identificar e a extrair informação linguística (lexical, sintáctica e semântica). Esta informação é utilizada como input de um modelo de tradução automático, construído previamente por algoritmos de aprendizagem supervisionada, obtendo-se como output uma proposta de tradução na Língua "destino".

Tal como referido, antes de se poder efectuar o processo de tradução automático é necessário realizar o processo de aprendizagem supervisionada, de forma a ser obtido o modelo de tradução a aplicar na segunda fase do processamento referido na figura 1.

Para o processo de aprendizagem supervisionada é utilizada a seguinte metodologia:

(1) são aplicadas as ferramentas de processamento de Língua Natural referidas anteriormente (que serão descritas na próxima secção) sobre um conjunto de textos existentes nas Línguas "origem" e "destino";
(2) a informação identificada e extraída é representada através de estruturas de dados típicas – grafos, árvores, listas – e é dado como input aos algoritmos de aprendizagem automática supervisionada, em conjunto com os pares de tradução Língua "origem" – Língua "destino" existentes no corpus de treino (ver secção "Aprendizagem Automática");
(3) o modelo obtido na alínea anterior é salvaguardado e será aplicado para a tradução de novos textos.

É de realçar que esta metodologia implica a necessidade de se ter um corpus paralelo nas duas Línguas, de dimensão razoável e que seja efectuado o seu alinhamento (de uma forma manual ou automática). Requer, ainda, que tenham sido tomadas decisões sobre a granularidade do alinhamento: seja a nível de frases, a nível de segmentos, a nível de entidades ou a nível de palavras.

Ferramentas para o Processamento da Língua Portuguesa

Embora a arquitectura apresentada na secção anterior seja genérica e independente da Língua, e como o foco deste trabalho é a tradução de/para a Língua Portuguesa, nesta secção são descritas algumas ferramentas computacionais existentes para o processamento da Língua Portuguesa, que poderão ser utilizadas para analisar os textos e para produzir como output informação linguística relevante para os algoritmos de aprendizagem automática.

- Etiquetadores morfo-sintácticos – PoS taggers

Existem vários etiquetadores morfo-sintácticos para a Língua Portuguesa disponíveis na web, com licenças de utilização típicas de software livre (GPL ou LGPL). Uma referência é o software "TreeTagger"[4], desenvolvido por Helmut Schmid (1994), que já foi aplicado a mais de 14 Línguas distintas, incluindo o Português. Uma referência alternativa, mais específica para a Língua Portuguesa, Galego e Castelhano, é o pacote de software FreeLing[5], que inclui várias ferramentas para o processamento de textos em Língua Natural, entre os quais etiquetadores morfo-sintácticos. É de realçar que este pacote de software também efectua um conjunto de operações prévias necessárias à esta tarefa: separação de frases, identificação de termos e análise morfológica, expansão de contracções. Para a tarefa de análise morfológica, o software Jspell[6] (Simões e Almeida, 2001) é também uma referência incontornável.

Exemplo: A frase "A Maria leu o livro." é analisada da seguinte forma pelo analisador FreeLing:

A o DA0FS0 0.667849

Maria maria NP00000 1

leu ler VMIS3S0 0.875

o o DA0MS0 0.944727

livro livro NCMS000 0.977273

. . Fp 1

- Reconhecimento de Entidades Nomeadas – NER Named Entity Recognition

O reconhecimento de entidades nomeadas (NER) é uma tarefa bastante relevante para o processo de tradução automática. Efectivamente, a identificação de, por exemplo, pessoas, entidades, locais e datas permitirá criar agrupamentos coerentes de termos e efectuar a sua análise de uma forma conjunta. O já referido pacote de software FreeLing incorpora módulos que permitem realizar esta tarefa com uma qualidade bastante razoável para a Língua Portuguesa. Um outro sistema disponível também para utilização é o Rembrandt[7] (Cardoso, 2008), que permite, ainda, a identificação de algumas relações entre as entidades identificadas.

Exemplo: O software FreeLing obtém o seguinte resultado para a frase "A Maria leu o livro em Lisboa.":

A o DA0FS0 0.667849

Maria maria NP00SP0 1

leu ler VMIS3S0 0.875

o o DA0MS0 0.944727

livro livro NCMS000 0.977273

em em SPS00 1

Lisboa lisboa NP00G00 1

4 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
5 http://nlp.lsi.upc.edu/freeling/
6 http://natura.di.uminho.pt/wiki/doku.php?id=ferramentas:jspell
7 http://xldb.di.fc.ul.pt/Rembrandt/

SIMPÓSIO 46

. . Fp 1

Realce-se a identificação da etiqueta “SP” (pessoa) no termo “Maria” e da etiqueta “G” (local) no termo “Lisboa.

- Analisadores sintácticos

A identificação e utilização de segmentos de frases no processo de tradução automática é uma das abordagens que apresenta resultados mais promissores actualmente. Os referidos segmentos poderão ter uma relação directa com o resultado da análise sintáctica das frases, sendo, por exemplo, os sintagmas nominais, preposicionais e verbais dessas frases. Para tal, é necessário ter a capacidade de efectuar uma análise sintáctica dos textos. Para a Língua Portuguesa, existem vários analisadores disponíveis: FreeLing (com um analisador de dependências e um analisador “superficial”[8]); DepPattern[9] (com um analisador de dependências). O analisador PALAVRAS[10] (Bick, 2000) é uma referência reconhecida na área da análise sintáctica da Língua Portuguesa; no entanto, está somente disponível para utilização via web, o que inviabiliza a disseminação da sua utilização.

Exemplos:
Os analisadores PALAVRAS e FreeLing obtêm os seguintes resultados para a frase “A Maria leu o livro”:
PALAVRAS:

```
SOURCE: Running text
1. a Maria leu o livro.
A1
UTT:cl(fcl)
.
|-S:g(np)
| |-D:pron(det “o” <artd> DET F S)        a
| |-H:prop(“Maria” F S) Maria
|-P:v(fin “ler” <fmc> PS 3S IND VFIN)     leu
|-Od:g(np)
  |-D:pron(det “o” <artd> DET M S)        o
  |-H:n(“livro” M S)    livro
```

FreeLing:

```
+sn_[
  espec-fs_[
    +j-fs_[
      +(A o DA0FS0 -)
    ]
  ]
  +grup-nom-fs_[
    +w-fs_[
      +(Maria maria NP00000 -)
    ]
  ]
  grup-verb_[
    +verb_[
      +(leu ler VMIS3S0 -)
    ]
]
sn_[
  espec-ms_[
    +j-ms_[
      +(o o DA0MS0 -)
    ]
  ]
  +grup-nom-ms_[
    +n-ms_[
      +(livro livro NCMS000 -)
    ]
  ]
]
F-term_[
  +(. . Fp -)
]
]
```

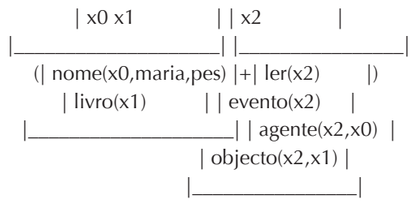
8 shallow parser
9 http://gramatica.usc.es/pln/tools/deppattern.html
10 http://beta.visl.sdu.dk/

SIMPÓSIO 46

- Análise semântica

Embora não seja viável efectuar uma análise semântica completa aos textos em Língua Portuguesa, é possível efectuar análises parciais, obtendo anotações semânticas ou, mesmo, estruturas de representação do discurso (DRS). Para o processo de inferência de anotações semânticas pode ser utilizado o software FreeLing; para as estruturas de representação do discurso, pode ser utilizada a ferramenta BOXER[11] (Bos, 2008), que transforma o resultado de análises sintácticas em DRS.

Exemplo:
A frase "A Maria leu o livro" pode ser transformada na seguinte DRS pela ferramenta BOXER (em formato gráfico e com algumas adaptações para a Língua Portugesa, para uma mais fácil visualização):



- Alinhadores de texto

- Conforme será referido na próxima secção, existem várias metodologias para a construção de sistemas de tradução automática baseados em métodos estatísticos, sendo que esses métodos pressupõem ou incorporam técnicas para o alinhamento automático de textos paralelos escritos em duas Línguas distintas. No entanto, existem alguns alinhadores de texto disponibilizados pela comunidade de investigadores de Língua Portuguesa que podem ser utilizados de uma forma autónoma: CEPRIL[12] e NATools[13].

Tradução Automática Estatística[14]

A tradução automática estatística utiliza técnicas de aprendizagem automática para abordar o problema da tradução de textos. Lopez (2008) efectua uma análise detalhada do estado da arte neste domínio e das várias técnicas utilizadas actualmente.

Uma análise cuidada das várias técnicas utilizadas leva-nos a concluir que o recurso a informação linguística tem vindo a ganhar uma importância crescente nos sistemas de tradução automática estatística, com o objectivo de melhorar quer os modelos das Línguas, que o modelo de tradução.

Vejamos alguns exemplos:

A) Modelos de tradução baseados em palavras
Os sistemas actuais baseados na equivalência entre palavras são tipicamente extensões aos modelos propostos pela IBM (Brown, 1990), com o objectivo de melhorar o seu desempenho. Entre algumas das extensões propostas inclui-se a utilização de informação complementar para cada palavra, incluindo etiquetas morfo-sintácticas, de forma a limitar (ou permitir) a reordenação das palavras.

B) Modelos de tradução baseados em segmentos
Nestes modelos, cuja relevância e desempenho tem vindo a aumentar nos últimos anos, a tradução é efectuada em grupos de palavras ou segmentos (Marcu, 2002). O sistema "open-source" Moses (Koehn 2007a) é um exemplo de uma ferramenta de tradução automática baseada em segmentos. Note-se que os segmentos podem ter (ou não) uma relação directa com as estruturas sintácticas das frases. No caso afirmativo (Wang, 2007), a importância de existirem ferramentas computacionais com a capacidade de obter estruturas sintácticas é fundamental. Outra abordagem possível à segmentação é a utilização das entidades nomeadas como identificador de segmentos com uma forte coerência interna e que devem ser analisados de uma forma atómica (Koehn, 2003).
Koehn (2007b; 2010) propôs uma extensão ao modelo baseado em segmentos, de forma a integrar informação linguística (ou outra) a nível das palavras – etiquetas morfo-sintácticas, lema, informação morfológica – tendo obtido melhorias nos resultados finais.

C) Modelos de tradução baseados em gramáticas sintácticas
A utilização de gramáticas, que incorporam conhecimento da Língua através do recurso a regras sintácticas, tem sido efectuada por diversos investigadores, com resultados bastante positivos (Wu, 1998; Yamada, 2001; Melamed 2004). As gramáticas de dependências também foram objecto de análise e aplicação neste domínio (Melamed, 2003).

D) Modelos da Língua "destino"
Para resolver ambiguidades no processo de tradução é fundamental modelar a Língua "destino". Este processo de modelação é tipicamente baseado na construção de modelos de "n-gramas" e no cálculo de probabilidades associadas à sequência de termos. No entanto, a incorporação de informação linguística mais profunda nestes modelos tem vindo a sofrer um incremento, através do recurso a modelos sintácticos (Wu, 1998; Marcu, 2006).

E) Processo de tradução (*decoding*)
O processo de tradução propriamente dito (*decoding*) é, em termos gerais, equivalente a um processo de análise de frases na Língua "origem" (*parsing*), sendo o resultado "lido" na árvore obtida. No entanto, e devido à existência de ambiguidades na Língua Natural e na sua representação, é possível que existam várias traduções possíveis para a mesma frase "origem". Nesta situação, é efectuada uma ordenação das propostas de tradução obtidas, com base no modelo da Língua "destino". Este modelo pode ser mais complexo e abrangente do que o que é utilizado no processo de tradução, incorporando modelos "n-gramas" e analisadores morfológicos e sintácticos. Uma abordagem alternativa é a utilização de Support-Vector Machines, tendo como *input* um conjunto de informação associada a cada frase (incluindo informação linguística), para classificar se as frases obtidas pertencem ou não à Língua "destino".

F) Modelos semânticos
A utilização de informação semântica no processo de tradução tem vindo a ser objecto de alguma atenção nos últimos anos, sendo uma área bastante promissora. No entanto, e devido à dificuldade em se realizarem análises semânticas dos textos, as abordagens existentes ainda são bastante preliminares, recorrendo tipicamente a informação semântica lexical (Chan, 2007) e não a uma análise semântica profunda.

Conclusões e Trabalho Futuro

Neste artigo foi analisada a importância do uso de informação linguística em modelos de tradução automática baseados em aprendizagem.

Em concreto, apresentou-se a arquitectura geral dos tradutores automáticos estatítisticos e discutiram-se as abordagens

11 http://svn.ask.it.usyd.edu.au/trac/candc/wiki/boxer
12 http://www2.lael.pucsp.br/corpora/alinhador/
13 http://linguateca.di.uminho.pt/natools/
14 Statistical Machine Translation

SIMPÓSIO 46

existentes e a importância da informação linguística nessas abordagens. Identificaram-se algumas das ferramentas computacionais existentes para a Língua Portuguesa, com licenças de software-livre, que poderão permitir a construção de tradutores automáticos optimizados para a Língua Portuguesa.

Como trabalho futuro, é fundamental aplicar estas metodologias e ferramentas computacionais a corpora de textos em Língua Portuguesa, para os quais existam textos paralelos noutras Línguas, de forma a avaliar as abordagens propostas e permitir a criação de sistemas "abertos" de tradução automática de/para a Língua Portuguesa.

Referências bibliográficas

Bick, Eckhard. 2000. The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework, *Aarhus University Press.*

Bos, Johan. 2008. Wide-Coverage Semantic Analysis with Boxer. In: J. Bos, R. Delmonte (eds): Semantics in Text Processing. STEP 2008 Conference Proceedings. Research in Computational Semantics. *College Publications. p. .277-286.*

Brown, P. F., Cocke, J., Pietra, S. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., Roossin, P. S. 1990. A statistical approach to machine translation. Computational Linguistics. *16, 2 (Jun), p. 79–85.*

Cardoso, Nuno. 2008. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto. Encontro do Segundo HAREM, PROPOR. *Aveiro. Portugal.*

Chan, Y. S., Ng, H. T., and Chiang, D. 2007. Word sense disambiguation improves statistical machine translation. In Proc. of ACL. *33–40.*

Gonçalves, Teresa; Quaresma, Paulo. 2008. Text classification using tree kernels and linguistic information. In: M. Arif Wani, Xue wen Chen, David Casasent, Lukasz Kurgan, Tony Hu, Khalid Hafeez (Eds). IMLA'08 – 7th International Conference on Machine Learning and Applications. *IEEE Computer Society. p. 763–768.*

Gonçalves, Teresa; Quaresma, Paulo. 2009. Using graph-kernels to represent semantic information in text classification. In: Petra Perner (Ed). Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Artificial Intelligence LNCS/LNAI. *Vol. 5632. Springer. p. 632–646.*

Gonçalves, Teresa; Quaresma, Paulo. 2010. Using linguistic information and Machine Learning Techniques to Identify Entities from Juridical Documents, In: E. Francesconi; E. Montemagni; W. Peters; D. Tiscornia (Eds). Semantic Processing of Legal Texts, Lecture Notes in Artificial Intelligence LNAI 6036, *Springer, p. 44-59.*

Koehn, Philipp; Haddow, Barry; Williams, Philip; Hoang, Hieu. 2010. More Linguistic Annotation for Statistical Machine Translation. *Fifth Workshop on Statistical Machine Translation and Metrics MATR.*

Koehn, Philipp; Hoang, Hieu; Birch, Alexandra; Callison-Burch, Chris; Federico, Marcello; Bertoldi, Nicola; Cowan, Brooke; Shen, Wade; Moran, Christine ; Zens, Richard; Dyer, Chris; Bojar, Ondrej; Constantin, Alexandra; Herbst, Evan. Moses: Open Source Toolkit for Statistical Machine Translation. *ACL 2007.* Demo.

Koehn, Philipp; Knight, Kevin. 2003. Feature-Rich Statistical Translation of Noun Phrases. *ACL.*

Koehn, Philipp; Hoang, Hieu. 2007b. Factored Translation Models, *EMNLP.*

Leal, Ana Luísa Varani. 2009. AUTEMA-DIS - Arquitetura Computacional para Identificação da Temática Discursiva em Textos em Língua Portuguesa. Tese de Doutoramento. Universidade de Évora. Portugal. 217 p.

Lopez, Adam. 2008. Statistical Machine Translation. ACM Computing Surveys. Volume 40, nº 3.

Marcu, D.; Wong, W. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. of EMNLP*. 133–139.

Marcu, D., Wang, W., Echihabi, A., Knight, K. 2006. SPMT: Statistical machine translation with syntactified target language phrases. In *Proc. of EMNLP*. 44–52.

Melamed, I. D. 2003. Multitext grammars and synchronous parsers. In *Proc. of HLT-NAACL.* 79–86.

Melamed, I. D. 2004. Algorithms for syntax-aware statistical machine translation. In Proc. Of *TMI*

.

Oliveira, Francisco; Wong, Fai; Hong, Iok-Sai; Dong, Ming-Chui. 2010. Parsing Extended Constraint Synchronous Grammar in Chinese-Portuguese Machine Translation. The International Conference on Computational Processing of Portuguese, former Workshop on Computational Processing of the Portuguese Language (PROPOR). *Porto Alegre. Brasil.*

Schimd, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. International Conference on New Methods in Language Processing. *Manchester. UK.*

Simões, Alberto Manuel; Almeida, José João. 2001. jspell.pm — um módulo de análise morfológica para uso em processamento de linguagem natural. In Actas da Associação Portuguesa de Linguística, p. 485–495

Wang, C., Collins, M., and Koehn, P. 2007. Chinese syntactic reordering for statistical machine translation. In Proc. of EMNLP-CoNLL. *737–745.*

Wu, D.; Wong, H. 1998. Machine translation with a stochastic grammatical channel. In Proc. of ACL-COLING. *1408–1415.*

Yamada, K.; Knight, K. 2001. A syntax-based statistical translation model. In Proc. Of ACL-EACL.

A TRADUÇÃO DE DEMONSTRATIVOS ANAFÓRICOS NO PAR LINGUÍSTICO CHINÊS-PORTUGUÊS: UMA ABORDAGEM DE BASE EM CÓRPUS

Marco ROCHA[15]

RESUMO: A investigação apresentada é uma abordagem de base em córpus para o estudo comparativo de demonstrativos anafóricos (DA) em português e em chinês, inclusive das estratégias de tradução conforme observadas através da análise de um córpus paralelo. Por demonstrativo anafórico, entende-se o demonstrativo usado como pronome independente, o que exclui as ocorrências como determinantes. Encontra-se em andamento a compilação de um córpus paralelo coletado manualmente, composto por textos bilíngues encontrados na Web. Espera-se atingir aproximadamente duzentas mil ocorrências nos originais em português, e uma quantidade equivalente nos originais chineses, tomando como base uma proporção palavra-caracter de 1:1,6. Neste momento, a investigação busca delimitar os padrões típicos de uso destes termos anafóricos em cada uma das línguas, através da análise de ocorrências em dois córpus monolíngues. Foram analisadas 110 ocorrências de DAs em chinês, extraídas do Lancaster Corpus of Mandarin Chinese, e 110 ocorrências em português brasileiro (*isto*, *isso*, e *aquilo*, incluindo contrações), extraídas do córpus Lácio-Ref. Esta amostra serve de referência para o desenvolvimento do estudo translinguístico.

PALAVRAS-CHAVE: fenômenos anafóricos; anotação de fenômenos discursivos; estudos da tradução de base em córpus; processamento computacional de línguas humanas.

1. Introdução

A pesquisa descrita em seguida visa a realizar uma investigação linguística contrastiva de base em córpus[16], a fim de estudar os demonstrativos anafóricos (doravante, DAs) em português e chinês. O demonstrativo anafórico típico, foco desta investigação, é exemplificado abaixo por uma unidade alinhada extraída do córpus paralelo português-chinês em construção. O negrito assinala o termo anafórico em questão.

> (1a) O tesouro está nas Pirâmides e **isto** você já sabia; mas teve que pagar seis ovelhas porque eu lhe ajudei a tomar uma decisão.
> (1b) 财宝就在金字塔附近，这你已经知道了。不过你得拿出六只羊做报酬。因为我帮你作出了一个决定。”

No presente estágio da pesquisa, procura-se estabelecer os padrões de uso dos DAs em cada uma das línguas através da análise de dados de dois córpus monolíngues. Os dados dos DAs para o português foram extraídos do córpus Lácio-Ref, compilado pelo Núcleo Interdisciplinar de Linguística Computacional (NILC). O córpus encontra-se disponível em http://www.nilc.icmc.usp.br/nilc/. Os dados do chinês foram coletados do Lancaster Corpus of Mandarin Chinese (doravante, LCMC), compilado por Anthony McEnery e Richard Xiao, disponível através do Oxford Text Archive (ota.ahds.ac.uk).

Os propósitos da investigação são essencialmente linguísticos e partem da convicção de que os estudos plurilinguísticos podem levar a resultados científicos que não poderiam ser obtidos em estudos monolíngues (Johansson 2007). A isso se soma a insuficiência de investigações a respeito das relações anafóricas com base em dados de uso real, sobretudo em relação aos DAs. Contudo, espera-se que os resultados obtidos possam ser úteis também para fins de tradução de máquina.

É provavelmente correto afirmar que, desde a publicação de Halliday and Hasan (1976), as investigações relacionadas a fenômenos anafóricos vêm crescendo sistematicamente em número e amplitude. Segundo Fox (1996), houve “uma explosão de pesquisa” sobre anáfora durante os anos oitenta. A partir de 1996, o tema de pesquisa passou a ter um evento bianual exclusivo, o *Discourse Anaphora and Anaphora Resolution Colloquium* (DAARC). Não obstante, os DAs não têm sido investigados na mesma medida que as outras formas de anáfora, provavelmente devido às variações bastante amplas na definição de anáfora.

Conforme apontado por Chu (1998), o termo anáfora pode ser compreendido tanto em um sentido restrito, que limita o termo à correferência entre grupos nominais e pronomes, quanto em um sentido amplo, incluindo qualquer elemento linguístico que se refira ao que já foi mencionado no texto. O sentido restrito deixaria de fora todas as formas de referência a passagens de discurso, as quais são bastante comuns através de DAs. A falta de consenso no que diz respeito às relações textuais, no âmbito das investigações linguísticas, reforça a tendência a desconsiderar os DAs.

Uma outra perspectiva importante nos estudos sobre anáfora diz respeito ao que se convencionou chamar de resolução de anáforas na terminologia relacionada à tecnologia das línguas humanas. Desde os estágios iniciais da pesquisa na área, percebeu-se que a identificação de um antecedente para um termo considerado como anafórico era uma tarefa difícil para sistemas de computadores que processam línguas humanas. Muitos avanços ocorreram desde que as relações anafóricas foram inicialmente trabalhadas em sistemas de computador nos anos setenta (Winograd 1972; Hobbs 1978). A carência de investigações focadas na resolução de demonstrativos anafóricos em sistemas de computador permance ainda assim real (ver, porém, Webber 1991; Byron 1998; Gundel, Hegarty e Borthen, 2001). No que diz respeito também à tradução automática, o processamento adequado das relações anafóricas, cuja necessidade foi enfatizada em Mitkov, Choi e Sharp (1996), continua sendo um desafio para os estudiosos da área. A investigação das relações anafóricas também se estendeu a estudos específicos sobre a língua chinesa, como Teng (1981) e Wu (1992). Mais recentemente, Huang (2000) publicou um estudo comparativo que inclui com destaque o chinês. Xu (2002) discute as anáforas nominais em textos chineses, e Hu (2008) apresentou tese de doutoramento sobre a resolução de anáforas zero do chinês em sistemas de computador.

A abordagem usada neste estudo tem como base a linguística de córpus, a qual pressupõe o uso de uma fonte de dados linguísticos, o córpus, da qual são extraídos exemplos de texto em situações reais, a fim de coletar informações para a elaboração de teorias que expliquem adequadamente os fatos observados da língua. Estas propostas explicativas muitas vezes assumem a forma de modelos para o desenvolvimento de aplicações de natureza linguística, tais como dicionários, gramáticas ou metodologias para o ensino de línguas. No campo da tecnologia das línguas humanas, as abordagens de base em córpus tornaram-se um paradigma de pesquisa de alta relevância, tanto usadas com a finalidade de dar sustentação ao processamento linguístico de base estatística, quanto, simplesmente, como uma fonte de dados para o aperfeiçoamento de sistemas com capacidade de processar línguas humanas.

O mesmo pode ser dito em relação a desenvolvimentos recentes em estudos da tradução. As abordagens de base em córpus tornaram-se rapidamente uma forma importante de investigação, logo que os pesquisadores tomaram consciência de que o uso de um córpus paralelo, isto é, um córpus que contém textos em uma língua dada e suas traduções para outra língua, constitui um instrumento poderoso para a análise contrastiva e plurilinguística (ver McEnery e Xiao 2007). O artigo supracitado aborda também o uso de córpus comparáveis, os quais não são constituídos por originais e suas traduções, mas por textos em duas ou mais línguas dentro de uma mesma estrutura de amostragem.

O restante deste artigo está organizado da seguinte forma: a próxima seção apresenta a metodologia considerada apropriada para a realização dos objetivos pretendidos; a terceira seção apresenta os resultados obtidos; a quarta seção apresenta conclusões e possíveis desenvolvimentos futuros.

2. Metodologia

A definição dos córpus usados na investigação é o primeiro aspecto metodológico abordado. A versão do Lácio-Ref usada contém 4.983.645 palavras, conforme a contagem realizada através do Recurso WordList do programa

15 UFSC, Centro de Comunicação e Expressão, Departamento de Língua e Literatura Vernáculas, Câmpus da Trindade, 88040 Florianópolis, SC, Brasil. marcor@cce.ufsc.br
16 A forma aportuguesada córpus será usada neste texto, inclusive para o plural.

WordSmith Tools, versão 5.0. O LCMC contém 841.400 palavras, ainda conforme a mesma ferramenta. As decisões quanto ao que constitui uma palavra em chinês não são triviais. Não obstante, os pesquisadores responsáveis pela compilação do LCMC (Xiao e McEnery) realizaram revisão cuidadosa do material incluído no córpus, que pode, portanto, ser considerado confiável.

O propósito desta análise de DAs em corpus monolíngües é detectar os padrões de uso destes termos anafóricos, inclusive a sua relação com os antecedentes. Os DAs foram classificados segundo a sua função sintática. Foram utilizadas quatro categorias, especificadas em seguida, a saber: sujeito de verbo copular; sujeito de verbo lexical; objeto de verbo; e objeto de preposição. O significado dos termos é semelhante ao utilizado em trabalhos de referência das línguas ocidentais, como gramáticas e similares. Abaixo são mostrados, em ambas as línguas, exemplos de casos de DA (destacados em negrito) como sujeito de verbo copular, retirados dos córpus mencionados acima:[17]

(2) Na região metropolitana, cerca de 40% do esgoto produzido é tratado. **Isso** é pouco, mas sem dúvida um avanço se comparado aos números de dez anos atrás, quando se tratava apenas 10%.
(3) 有 几 位 先生 谈到 中学 语法 教学
Yǒu jǐ wèi xiānshēng tándào zhōngxué yǔfǎ jiàoxué
teve diverso (cl.) professor falar sobre ensino médio gramática ensinar
中 遇到 的 问题， 这 也 是 大家
zhōng yùdào de wèntí, zhè yě shì dàjiā
em encontrar (de) problema, isso também é todo mundo
都 很 关心 的 问题。
dōu hěn guānxīn de wèntí.
todos muito preocupar (de) problema.

"Houve alguns professores que falaram sobre problemas encontrados no ensino de gramática do ensino médio, isso também é um problema com que todos se preocupam muito."

Abaixo são apresentados exemplos de casos classificados como sujeitos de verbos lexicais em ambas as línguas:

(4) São indivíduos que muitas vezes ficam um, dois, três anos, o resto da vida em tratamento para poder se recuperar do acidente e precisam de uma equipe de atendimento altamente especializada. **Isso** sai do bolso de todos nós.
(5) 在 李 懋 的 眼里， 职位 的 不断 上升，
Zà i li mào de yǎn lǐ, zhíwèi de bùduàn shàngshēng,
Em Li Mao (de) olhos-dentro, posição (de) constante ascender,
就 意味 着 社会 地位 和 社会 价值
jiù yìwèi zhe shèhuì dìwèi hé shèhuì jiàzhí
exatamente significado (zhe) sociedade posição e sociedade valor
的 不断 提升，这 使得 李 懋 感到 极其 充实，
de bùduàn tíshēng, zhè shǐde li mào gǎndào jíqí chōngshí,
(de) constante melhorar, isso fazer Li Mao sentir muito fortalecer,
极其 地 安全。
jíqí de ānquán.
muito (de) segura.
"Na opinião de Li Mao, a ascensão constante na carreira significa que sua posição e status social melhoram constantemente. Isso faz Li Mao sentir-se muito fortalecida, muito segura."

17 Os exemplos em chinês são apresentados em alinhamento com a transcrição em pinyin, na linha subsequente, e uma glosa em português. No final, é apresentada uma tradução global sem alinhamento com o original. As partículas do chinês são apenas repetidas entre parênteses na glosa, e o código (cl.) é usado para os classificadores.

São apresentados em seguida ocorrências extraídas do Lácio-Ref para exemplificar as categorias objeto de verbo (6) e objeto de preposição (7).

(6) No caso particular do ABN Amro, representou também a possibilidade de emitir US$ 65 milhões em eurobônus. Com o capital anterior, de cerca de US$ 100 milhões, o banco não poderia fazer **isso**.
(7) "Pesquisamos com os alunos as origens das danças e dos instrumentos. Para **isso**, eles entrevistam os moradores mais velhos", conta Maria Eunice dos Santos, monitora do grupo do siriri.

No que diz respeito aos antecedentes, foram utilizadas duas variáveis dicotômicas para realizar a sua classificação. A primeira deles distingue os antecedentes explícitos, aqueles previamente presentes no texto (exemplo (8) abaixo), dos antecedentes implícitos (exemplo (9) abaixo), os quais têm que ser inferidos a partir de informações presentes no texto. Os exemplos servem também para demonstrar que a decisão de atribuir cada ocorrência a uma das categorias não é trivial e pode gerar bastante "agonia analítica".

(8) A proporção de trabalhadores com carteira, em fevereiro de 2002 (40,4%) era menor do que a do mesmo mês do ano anterior (40,8%) e pouco maior do que a de 2000 (40,1%). **Isso** significa que não há ainda evidências de que se tenha reduzido a informalidade na RMSP, em função do crescimento do emprego formal.
(9) O IQSC é um instituto que se firmou principalmente em termos de pesquisa. Sua pós-graduação é bastante conhecida e, agora, realmente tem a necessidade de se expandir. Atualmente, formamos químicos, bacharéis, tecnólogos e temos que estender **isso** a outras habilitações.

No exemplo (9), o antecedente implícito do DA é *esta formação*, sobre a qual se falava anteriormente. Além de extrair o grupo nominal do verbo da oração coordenada precedente, o leitor ou ouvinte precisa compreender o termo como uma referência às características do instituto mencionadas anteriormente. Pode-se concluir, desse modo, que o antecedente não está explicitamente presente no texto.

A segunda variável separa os antecedentes nominais dos antecedentes textuais. A primeira categoria classifica os antecedentes que podem ser expressos por um grupo nominal (exemplo (10) abaixo), enquanto a segunda se aplica aos antecedentes expressos por um segmento de texto (exemplo (11)). Mais uma vez, a decisão analítica pode ser bastante difícil no caso dos antecedentes implícitos.

(10) Como os dados relativos ao mutante snf1 não são confiáveis, nota-se que, de uma forma geral, há uma diminuição no valor do fluxo pela via PP (Tabela 4.11). **Isto** ocorre provavelmente devido à composição da biomassa dos mutantes, a qual assume ... da coenzima NADPH para biossíntese de aminoácidos.
(11) A entropia é a segunda lei do calor. Ela diz que é impossível a transmissão de calor do mais fraco para o mais forte. A Terra recebe calor do Sol, e ele jamais vai retirar esse calor da Terra. **Isso** não acontece só com os astros, acontece com todos os elementos.

Como pode ser observado nos exemplos acima, tanto os antecedentes nominais quanto os textuais exigem procedimentos de identificação de complexidade variável. O processo de classificação das ocorrências da amostra foi em consequência expandido para abranger um comentário sobre a posição do antecedente em relação ao DA. No caso do exemplo (8), por exemplo, o antecedente é a sentença anterior plena. No caso do exemplo (10), o objeto do verbo da oração subordinada objetiva direta precedente é o antecedente, em um procedimento de identificação que enfatiza o fator de proximidade. Já no exemplo (11), o antecedente é a oração subordinada objetiva direta que integra a sentença anterior à precedente. A classificação procura incorporar estes elementos nos referidos comentários, de modo a estabelecer estratégias de identificação de antecedentes associadas à tipologia geral de DAs e antecedentes. Não foram utilizadas categorias previamente definidas devido às incertezas ainda existentes em relação à sistematização da variabilidade destes posicionamentos.

SIMPÓSIO 46

A metodologia de coleta teve como ponto de partida a existência de DAs em português que não ocorrem como determinantes. São estes: *isto*, *isso*, *aquilo* e suas respectivas contrações com as preposições *de* e *em*. Foram geradas concordâncias para as nove formas dos DAs, com a subsequente determinação dos padrões mais importantes com base sobretudo na frequência de coocorrência. Em chinês, de forma semelhante, a coleta teve como alvo os demonstrativos 这 e 那. Porém, estes demonstrativos não ocorrem apenas como termos anafóricos. Foi necessário, deste modo, analisar cada uma das ocorrências das concordâncias para determinar quais eram de fato casos de anáfora. Em seguida, os padrões dentro dos casos de anáfora foram estabelecidos da mesma forma. Finalmente, uma amostra de 110 casos em cada uma das línguas foi analisada com uso das variáveis descritas anteriormente. Os 110 casos foram coletados com base em um critério de proporcionalidade em conformidade com as proporções de cada um dos padrões encontrados em relação ao total de número de casos de DAs em cada uma das línguas. O procedimento deve ficar claro após a especificação das frequências dos padrões para os córpus como um todo que abre a seção de resultados a seguir.

4. Resultados

Dentre os DAs da língua portuguesa investigados, o pronome neutro *isso* é o mais frequente. Sem considerar as contrações, em um total de 6699 de DAs no Lácio-Ref, há 4.946 ocorrências deste pronome, isto é, 73,83%. Há 1467 ocorrências de *isto* (21,90%), enquanto o distal neutro *aquilo* chega apenas a 286 ocorrências (4,27%). Dentre as contrações, as discrepâncias de frequência são ainda mais acentuadas, com um predomínio da contração *disso*, a qual, com 1353 ocorrências em um total de 1671 contrações, constitui 80,96% da amostra. As demais frequências são: *nisso*, com 85 ocorrências (5,08%); *disto*, com 100 ocorrências (5,98%); *nisto*, somente 14 ocorrências (0,83%); *daquilo*, 91 ocorrências (5,44%); e *naquilo*, com apenas 28 ocorrências (1,67%). As frequências somadas dos DAs analisados e suas contrações chegam a 8370 ocorrências, ou seja, 0,16% do total de palavras do córpus. A Tabela 1 abaixo apresenta as frequências absolutas e relativas dos principais padrões de ocorrência do DA *isso* no Lácio-Ref. Os percentuais dizem respeito ao total de ocorrências do pronome *isso*.

Tabela 1 – Padrões de ocorrência do DA *isso*

por isso	959	19,38%
para isso	408	8,24%
isso é	404	8,16%
com isso	390	7,88%
isso não	317	6,41%
que isso	279	5,64%
tudo isso	200	4,04%

Os dois padrões mais frequentes, como também o quarto mais frequente, tornam clara a importância do DA como parte integrante de locuções conjuntivas, muitas vezes seguidas de vírgula e precedidas da conjunção **e**. O padrão *que isso* aponta para a ocorrência do DA como sujeito de orações objetivas diretas, nas quais *que* é a conjunção integrante ligada a verbos da oração principal como os verbos *achar*, *acreditar*, *alegar*, *concordar*, *pensar* e *saber*. O padrão também ocorre em orações subjetivas ligadas a formas do verbo *ser*, seguidas de adjetivo (*é claro que isso...*; *é possível que isso...*), ou a outros verbos de ligação (*parece que isso...*). Também está dentro deste padrão, com 29 ocorrências, a colocação *para que isso*.

Dentro do padrão *isso não*, há 56 ocorrências de *isso não é*, as quais, se somadas ao padrão com o verbo afirmativo, tornariam o grupo o segundo mais frequente do total de ocorrências. Outro agrupamento numeroso dentro do padrão em questão é *isso não significa*, com 31 ocorrências, as quais também poderiam ser somadas às 86 ocorrências de *isso significa*. O padrão *tudo isso* (200 ocorrências) tem a característica de permitir uma referência abrangente a passagens de texto ou enumerações. Os sete padrões mais frequentes listados na Tabela 1 somam 2957 ocorrências, praticamente 60% do total de ocorrências de *isso*, sinalizando um alto grau de padronização. O efeito é acentuado por outros padrões frequentes abaixo das 200 ocorrências, como *fazer isso* (75 ocorrências); *isso ocorre* (43) e *isso acontece* (43). O fenômeno da padronização é ainda mais forte para a contração com a preposição *de*. Do total de 1353 ocorrências, os sete padrões mais frequentes somam 1029 ocorrências (76,05%), sendo que o padrão mais frequente, a locução conjuntiva *além disso*, constitui 60,53% deste total. A Tabela 2 abaixo resume os padrões da contração.

Tabela 2 – Padrões de ocorrência do DA *disso*

além disso	819	60,53%
apesar disso	60	4,43%
nada disso	39	2,88%
exemplo disso	35	2,58%
depois disso	35	2,58%
antes disso	20	1,47%
em vista disso	21	1,55%

A padronização acentuada persiste em relação aos dados para o DA *isto*. Das 1467 ocorrências, o padrão *isto é* perfaz 782 (53,30%). O padrão é sintática e semanticamente distinto do padrão *isso é* do DA *isso*. Na enorme maioria dos casos, a colocação aparece isolada por sinais de pontuação, sobretudo entre vírgulas. Sinaliza uma explicação relacionada ao texto anterior, ou ainda uma auto-correção. Há poucas exceções na amostra. A Tabela 3 abaixo apresenta os padrões para o DA *isto*.

Tabela 3 – Padrões de ocorrência do DA isto

isto é	782	53,30%
isto não	52	3,54%
por isto	37	2,52%
isto significa	36	2,45%
para isto	31	2,11%
com isto	22	1,49%

A contração *disto* apresenta os mesmos dois padrões predominantes da contração *disso*, ainda que as frequências sejam bem mais baixas. O total de 100 ocorrências contém 41 casos de *além disto* (41%) e 10 casos de *apesar disto* (10%). Percentualmente, o padrão com a preposição *além* é menos frequente, mas o padrão com a preposição *apesar* é mais frequente.

O DA distal *aquilo* também apresenta um padrão dominante, a sequência *aquilo que*, com 201 ocorrências (70,27%) das 286 registradas. O fenômeno se repete com maior intensidade nos dados das contrações *daquilo* e *naquilo*. Dentre as 91 ocorrências da primeira contração encontradas no córpus, 76 são constituídas pela sequência *daquilo que* (83,51%). Para a segunda contração, o percentual aproxima-se do número total de ocorrências, já que 25 das 28 ocorrências são da colocação com a palavra *que*.

A despeito do número bem mais alto de ocorrências em relação a *naquilo*, a contração *nisso* não apresenta padrões definidos com a mesma clareza. A sequência *nisso que* é a colocação mais frequente, mas com apenas 10 dos 85 casos coletados (11,76%). Vale destacar os oito casos da colocação *pensando nisso*, uma vez que a forma de gerúndio do verbo é muito mais rara, o que caracteriza uma função discursiva marcada que deve ser investigada com maior cuidado. A frequência muito baixa da contração *nisto* resulta em uma ausência de padrões claros.

Os DAs da língua chinesa exigem que a investigação utilize procedimentos bastante diferentes de análise de dados. Conforme mencionado anteriormente, os demonstrativos 这 e 那, assim como as formas de plural 这些 e 那些, não ocorrem apenas como termos anafóricos. A análise das concordâncias, visando a identificação de DAs, é muito mais trabalhosa, uma vez que os recursos computacionais existentes, mesmo em combinação com as etiquetas de classe de palavras do LCMC, não são capazes de identificar os pronomes demonstrativos independentes tipicamente anafóricos e separá-los dos determinantes.

O demonstrativo mais frequente é 这, com 4142 ocorrências (72,79% do total das formas singulares de demonstrativos). Em seguida, vem o demonstrativo 那, que ocorre 1548 vezes no córpus (27,21%). Estes percentuais dizem respeito à frequência somada destes dois DAs. A forma de plural 这些 aparece 672 vezes no LCMC. Já a forma de plural 那些 ocorre 221 vezes. A Tabela 4 apresenta os padrões mais frequentes para o demonstrativo 这.

Tabela 4 – Padrões de ocorrência de这

这 是	612	14,77%
这 一	492	11,87%
了 这	177	4,27%
这 位	166	4,00%
在 这	158	3,81%
这 两	109	2,63%
说 这	106	2,56%

O grau de padronização do uso de 这 é consideravelmente menor do que o observado para os DAs mais frequentes de português. Os sete padrões mais frequentes chegam a 43,94% do total de ocorrências. O fato não é surpreendente, tendo em vista a variedade bem maior de funções que o termo desempenha na língua chinesa, se comparada à especialização como pronome independente de *isso*, *isto* e *aquilo* e suas contrações em português. O único padrão claramente associado à referência anafórica como pronome independente é 这 是,a colocação mais frequente, à qual foram acrescentadas, para fins de amostragem, ocorrências aleatoriamente coletadas das sequências também características de presença de anáfora,: 这 就是,com 93 ocorrências no córpus; 这 不 是,com 41 ocorrências; e 这 也 是, com 25 ocorrências. Os demais casos exigem maior investigação, já que seu possível caráter anafórico não pode ser detectado através dos recursos computacionais típicos de análise de córpus.

O pronome 那 ocorre 1548 vezes no LCMC. De modo semelhante à forma proximal, os casos de anáfora estão associados à sequência 那 是, que ocorre 137 vezes, e às combinações 那 就是, 那 不 是 e 那 也 是. Porém, a colocação com 就 como caracter independente, a qual, no caso dos padrões associados a 这, ficou abaixo do ponto de corte para os padrões incluídos na amostra, assume maior importância percentual para os padrões associados a 那. Há casos da colocação que estão associados a ocorrências anafóricas, como no exemplo (12) abaixo:

(12) " 唔 ， 很 好 ， 可见 他们 将 秘密
“ Wú, hěn hǎo, kějiàn tāmen jiāng mìmì
Bem, muito bom, então eles (jiāng) segredo
保守 得 十分 周全 ， 如果 连
bǎoshǒu de shífēn zhōuquán, rúguǒ lián
guardar (de) extremamente completo, se incluir
你 也 知道 了 ，那 就
ní yě zhīdào le, nà jiù
você também saber (le), isso exatamente
不 算是 秘密 喽 ！ "
bù suànshì mìmì lou! “
não considerar-como segredo (lou)!

“Bom, muito bem, dá para ver que eles guardaram completamente o segredo. Se até você também ficou sabendo, isso não pode mais ser considerado como um segredo!

A Tabela 5 resume os números para os padrões associados a 那. Apenas cinco padrões foram considerados relevantes para a amostra. Diferentemente da tabela anterior, os percentuais se referem ao total de ocorrências dos cinco padrões selecionados para inclusão na amostra.

Tabela 5 – Padrões de ocorrência de那

那 是	137	61,17%
那 就	48	21,42%
那 就是	31	13,83%
那 不 是	6	2,68%
那 也 是	2	0,90%

Os padrões anafóricos associados ao verbo 是 não se repetem para o plural proximal 这些. A identificação de casos de anáfora só parece possível através da análise manual caso a caso. O mesmo foi observado em relação ao plural distal 那些.

A amostra de 110 ocorrências analisada segundo o modelo de classificação apresentado anteriormente foi proporcional aos percentuais de ocorrência de cada uma das nove formas de DA do português incluídas no estudo, tendo como referência o total de DAs e suas contrações. As proporções foram arredondadas e ligeiramente alteradas para permitir a inclusão de pelo menos um caso de cada uma das nove formas. Deste modo, a amostra contém 66 ocorrências de *isso*; 19 ocorrências de *isto*; 17 ocorrências de *disso*; 2 ocorrências de *disto*; 2 ocorrências de *daquilo*; 2 ocorrências de *nisso*; e uma ocorrência de cada uma das formas *nisto* e *naquilo*.

Dentro destes estratos proporcionais, a amostra foi coletada aleatoriamente, através do recurso de concordância aleatória do WordSmith 5.0. O impacto dos padrões associados às preposições é bastante claro na frequência predominante de objetos de preposição. Por outro lado, a ocorrência de DAs na função de objeto de verbo é bastante baixa. A diferença entre os dois tipos de sujeito é relativamente pequena, mas demonstra a importância do padrão associado ao verbo *ser*. A Tabela 6 abaixo apresenta as frequências absolutas e relativas – estas últimas arredondadas – encontradas na amostra segundo a classificação das funções sintáticas dos termos anafóricos:

Tabela 6
Frequências das funções sintáticas dos DAs em português

Sujeito de verbo copular	32 (29%)
Sujeito de verbo lexical	24 (21%)
Objeto de verbo	9 (8%)
Objeto de preposição	47 (42%)
Total	110 (100%)

Os antecedentes destas 110 ocorrências de demonstrativo anafórico foram analisados segundo as duas variáveis dicotômicas apresentadas anteriormente. Há um predomínio amplo dos antecedentes explícitos em relação aos implícitos, uma vez que há 104 antecedentes do primeiro tipo, contra apenas 6 casos da segunda categoria. Com número tão baixo, não seria possível detectar padrões para este tipo de antecedente. Não foi feito, em consequência, nenhum cruzamento com qualquer das outras variáveis. Já a distribuição das frequências de antecedentes nominais e textuais é mais equilibrada, com a existência de 83 antecedentes textuais (75,45%) e 27 nominais (24,55%). A Tabela 7 apresenta os resultados do cruzamento entre as funções sintáticas dos DAs e a classificação da estrutura dos antecedentes.

SIMPÓSIO 46

Tabela 7
Distribuição das funções sintáticas dos DAs por estrutura do antecedente

Categoria	Nominal	Textual	Total
Sujeito de verbo copular	11	21	32
Sujeito de verbo lexical	8	16	24
Objeto de verbo	4	5	9
Objeto de preposição	4	43	47
Total	27	83	110

A despeito da amostra relativamente pequena, há uma relação clara entre a função sintática e a estrutura do antecedente, uma vez que quase a totalidade dos DAs com função de objeto de preposição referem-se a antecedentes textuais. Além disso, nos quatro casos em que um DA que é objeto de preposição tem um antecedente nominal, o grupo preposicional modifica um grupo nominal, como um substantitvo ou um adjetivo, e não o verbo, como nos exemplos (13) e (14) abaixo :

(13) A meta de R$ 9 bilhões, com possibilidade de expansão extra de 5% **disso**, conforme constava da minuta da MP (medida provisória) divulgada pela Folha, passou para R$ 9,5 bilhões mais 20% extras na MP finalmente adotada.
(14) Em sociedades que enfatizam a importância da comunidade, o respeito pela imagem positiva, por uma pessoa com comportamento dentro **daquilo** que é previsto, reafirma o valor do grupo.

Parece possível, portanto, mapear os contextos sintáticos em que existe a possibilidade de um antecedente nominal para um DA na função de objeto de preposição, já que os casos de antecedente textual parecem concentrar-se em objetos de preposição que modificam o verbo ou têm função conjuntiva, tipicamente colocados no início da sentença. Também existem indicações de que o antecedente pode ser especificamente identificado nestes casos. Na amostra, os antecedentes são geralmente o próprio nome modificado pelo grupo preposicional, como no exemplo (14). A única exceção na amostra é o exemplo (13), onde o antecedente é o núcleo do grupo nominal modificado através de um aposto que contém o grupo preposicional do qual o DA é parte. Ainda assim, a estrutura apositiva também pode ser mapeada segundo parâmetros puramente sintáticos e, portanto, viáveis para um processamento automático.
Os demais antecedentes nominais estão distribuídos pelas outras três categorias que classificam a função sintática do DA. No caso dos objetos de verbo, parece possível identificar os verbos que preferem antecedentes textuais, tipicamente formas dos verbos *fazer*, *decorrer* e *discutir*, ainda que a expansão da amostra deve seguramente levar a variação neste padrão, obrigando uma maior definição das características do padrão. Os DAs com função de objeto de verbo e antecedentes nominais, na amostra, aparecem como objetos de formas do verbo *pensar* em referência catafórica, como no exemplo (15) abaixo.

(15) Uma vez assumindo as funções de governo, eu tenho certeza de que nós todos estaremos pensando somente **nisso**: no interesse do país e no interesse da população.

Os DAs com antecedente nominal que ocorrem na função de sujeito de verbo copular são todos ocorrências do padrão *isto é*, exceto por um caso. Já em relação aos DAs nesta função com antecedentes textuais, há apenas três casos de ocorrência do padrão. Nestes três casos, a colocação **isto é** tem uma função conjuntiva, ligando orações, como no exemplo (16) abaixo, enquanto no exemplo (17), típico dos antecedentes nominais, um grupo nominal torna mais claro o significado do antecedente também nominal:

(16) Desse modo, ele extraiu, com toda clareza, dos próprios fatos, o que até então não fizera senão deduzir, semi-aprioristicamente, de materiais insuficientes, **isto é**, que a crise do comércio mundial, ocorrida em 1847, fôra a verdadeira mãe das revoluções de fevereiro e de março...
(17) Quando as instituições financeiras sabem que não são adequadamente monitoradas pelos depositantes **(isto é**, seus credores), elas têm incentivos em tomar maiores riscos com os seus depósitos.

A tendência apresenta variações, ainda que raras, exigindo maiores investigações para um estabelecimento de padrão. No que diz respeito aos sujeitos de verbos lexicais, não foi ainda detectado nenhum padrão ou mesmo tendência.
A amostra de DAs do chinês tem características bastante distintas, conforme mencionado anteriormente. Na amostra coletada, foram incluídos 80 casos de 这 e 30 casos de 那, preservando a proporcionalidade do total de ocorrências. Tendo em vista o predomínio quase absoluto de ocorrências com função de sujeito de verbo copular (104 ocorrências, com apenas 3 ocorrências de sujeito de verbo lexical; 3 ocorrências de objeto de verbo; e nenhuma ocorrência de objeto de preposição), o cruzamento desta variável com as demais não produz resultados úteis. As duas variáveis de classificação dos antecedentes, porém, apresentam proporcionalidade distinta da encontrada em português.

Há 20 casos de antecedentes implícitos (18,18%) na amostra de DAs do LCMC, uma proporção bem maior do que os apenas 4 da amostra do Lácio-Ref (3,63%). É provável que isto seja consequência da estrutura de amostragem de cada um dos córpus, que difere no que diz respeito à presença de textos ficcionais, de frequência relativa bem mais alta no LCMC. Uma vez que os textos ficcionais contêm diálogos em que ocorrem referências dêiticas ao ambiente em que transcorre a história, a posssibilidade de ocorrerem antecedentes implícitos na situação aumenta, como no exemplo (18) abaixo.

(18) B 科长 出示 了 搜查证 。 " 这 是 教廷 ，
B kē zhǎng chūshì le sōuchá zhèng. “ Zhè shì jiàotíng,
B divisão chefe mostrar (le) mandado de busca. “ Isso é Santa Sé.
不许 你们 侵犯 教廷 的 尊严 ！ "
bùxǔ nǐmen qīnfàn jiàotíng de zūnyán! “
Não-poder vocês violar Santa Sé (de) dignidade!

“O chefe de divisão B mostrou um mandado de busca. “Aqui é a Santa Sé. Vocês não podem violar a dignidade da Santa Sé.””

O DA refere-se ao local onde se encontram os personagens da história. O predicativo do sujeito serve como sinal para a identificação do antecedente da referência dêitica, já que há ocorrências anteriores do mesmo termo 教廷, modificado pelo locativo 罗马 (Roma). Nesta ocorrência, o nome comum refere-se ao território e à autoridade do Vaticano sem necessidade de repetir o locativo. Esta forma de referência dêitica, cuja resolução é sinalizada pelo predicativo, é frequente entre os DAs com antecedente implícito.

As proporções também diferem em relação à classificação da estrutura do antecedente. Há 39 antecedentes nominais (35,45%), aproximadamente dez pontos percentuais a mais do que na amostra da língua portuguesa. As referências dêiticas descritas acima contribuem para este aumento. Um outro fator que influencia no aumento de antecedentes nominais é o uso do pronome em construções coordenadas assindéticas que geralmente teriam a forma de orações relativas na língua portuguesa, as quais não utilizam DA.

Finalmente, o comentário relativo à posição dos antecedentes, acrescentado à anotação das ocorrências com base nas variáveis já descritas, ainda não pode ser considerado uma variável, uma vez que a variedade de situações obriga a uma subcategorização que exige uma amostra muito maior para ter alguma eficácia. Não obstante, há um grau de regularidade que pode se provar bastante útil na definição de padrões de ocorrência dos DAs, sobretudo na identificação de seus antecedentes. A Tabela 8 abaixo apresenta os primeiros resultados desta análise no que diz

respeito aos antecedentes textuais em português. Os dois casos de antecedentes implícitos considerados textuais são, como seria de se esperar, uma avaliação do analista, como também a base de inferência especificada.

Tabela 8
Posição dos antecedentes textuais

Sentença anterior	47 (56,62%)
Oração coordenada anterior	14 (16,86%)
Oração subordinada da sentença anterior	12 (14,45%)
Oração subordinada da própria sentença	3 (3,61%)
Sentença precedente à anterior	4 (4,81%)
Catáfora: oração subsequente	1 (1,20%)
Implícito: inferência de texto de quatro orações	1 (1,20%)
Implícito: inferência de texto de duas orações	1 (1,20%)
Total	83 (100%)

Como é possível observar, existe um grau de padronização que pode ser sistematizado a partir de uma amostra maior. A identificação dos antecedentes textuais pode ser tratada como uma variável incorporada ao modelo, embora seja necessário aprofundar a categorização e expandir a amostra. De todo modo, é possível generalizar que o antecedente típico de um DA anafórico do grupo analisado é um antecedente explícito textual expresso pela sentença anterior àquela em que o DA ocorre. Esta especificação cobre 42,72% do total de 110 ocorrências incluídas na amostra. A tabela 9 apresenta os resultados do comentário relativo ao posicionamento dos antecedentes nominais.

Tabela 9
Posição dos antecedentes nominais

Objeto de verbo da sentença anterior	12 (42,72%)
Objeto de preposição da sentença anterior	5 (18,51%)
Sujeito da sentença anterior	3 (11,11%)
Implícito: nominalização de verbo anterior	2 (7,40%)
Implícito: gr. nom. inferido de texto anterior	2 (7,40%)
Catáfora: grupo nominal subsequente	1 (3,70%)
Objeto da oração principal	1 (3,70%)
"isso tudo": dois parágrafos anteriores	1 (3,70%)
Total	27 (100%)

O número de antecedentes nominais é pequeno, o que dificulta a visualização de tendências. Contudo, é possível afirmar que o antecedente nominal típico de um DA é o objeto da sentença anterior. A análise do posicionamento dos antecedentes da amostra de DAs em chinês ainda precisa ser melhor desenvolvida e não será abordada aqui. A principal dificuldade em relação a este tipo de análise em chinês reside na necessidade de adaptar substancialmente a ideia de um posicionamento sintático do antecedente em relação a um termo anafórico à estrutura sintática da língua chinesa.

5. Conclusões e desenvolvimentos futuros

Este trabalho descreve o primeiro estágio de pesquisa em andamento, no qual uma amostra de 110 ocorrências de DAs em cada uma das duas línguas incluídas no projeto de pesquisa foi analisada com o intuito de estabelecer os padrões característicos de uso a partir de dois córpus monolíngues. O próximo estágio será concluir a análise das ocorrências de DAs em um córpus paralelo. Concomitantemente, a análise deste material monolíngue será utilizada para aperfeiçoar a especificação dos padrões encontrados, sobretudo no que diz respeito ao aperfeiçoamento da categorização relativa ao posicionamento dos antecedentes. Espera-se que a associação desta categorização às demais variáveis e a outros elementos léxico-gramaticais ainda a serem descobertos permita a organização de uma base de conhecimentos útil para módulos de resolução de anáforas em cada uma das línguas e para programas de tradução automática.

A adaptação da categorização de posicionamento do antecedente à língua chinesa ainda se encontra particularmente carente de aperfeiçoamento, tendo em vista características específicas da sintaxe chinesa, já que há diferenças importantes nas exigências de classificação, mesmo no que diz respeito ao nível de análise hierarquicamente anterior, as classes de palavras. Como quase sempre é verdade, a expansão da amostra é também um aspecto importante da continuidade da investigação.

6. Agradecimentos

A pesquisa em andamento cujos estágios iniciais foram descritos neste texto só é possível em consequência da concessão da bolsa BEX 6566-10-3 pela Coordenação de Aperfeiçoamento Pessoal do Ensino Superior (CAPES). Fico grato também a Luo Gang (Universidade de Zhejiang), Long Manying (Guilin University of Electronic Technology) e Ye Li (Universidade Federal de Santa Catarina) pela inestimável ajuda com a análise dos textos extraídos do córpus de chinês.

Referências bibliográficas

Byron, D.K. and Allen, J.F. 1998. Resolving demonstrative anaphora in the TRAINS93 corpus. In: *Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*, pps. 68-81. University of Lancaster.

Chu, Chauncey. 1998. *A discourse grammar of Mandarin Chinese*. New York: Peter Lang.

Fox, B. 1996. *Studies in anaphora*. Amsterdam/Philadelphia: John Benjamin.

Halliday. M.K. and Hasan, R. 1976. *Cohesion in English*. London: Longman.

Hobbs, J. 1986. Resolving pronoun references. In: Webber, B., Grosz, B. and Jones, K. (eds.), *Readings in natural language processing*. Palo Alto, CA: Morgan Kaufmann.

Huang, Y. (2000). *Anaphora: a cross-linguistic study*. Oxford: Oxford University Press.

Hu, Quinan (2008). *A corpus-based study on zero anaphora resolution in Chinese discourse*. Doctoral dissertation, City University of Hong Kong.

Johansson, S. 2007. Seeing through multilingual corpus. In: Facchinetti, R. (ed.), *Corpus linguistics 25 years on*. Amsterdam: Rodopi.

McEnery, Tony and Richard Xiao (2007). Parallel and comparable corpus: The state of play. In Y. Kawaguchi, T. Takagaki, N. Tomimori and Y. Tsuruga (eds.). *Corpus-Based Perspectives in Linguistics*. Amsterdam: John Benjamins. 131–145.

Mitkov, Ruslan, Sung-Kwon Choi, S.K. and Sharp, R. (1995). Anaphora resolution in machine translation . In: *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95)*, Leuven, Belgium.

Teng shou-hsin (1981). Deixis, anaphora and demonstratives in Chinese. *Cahiers de linguistique - Asie orientale*, Volume 10, Numéro 1, pp. 5-18.

Webber, B. (1991). Structure and ostension in the interpretation of discourse deixis. In: *Language and cognitive processes*, 6(2), pp.107-135, Psychology Press.

Wu, Guobin (1992). Discourse anaphora in Chinese: a rethorical predicate account. In: Harlow, S.J. and Warner, A.R. (eds.), *York papers in Linguistics*, 16, 185-202.
Xu, Jiujiu (2002). *Noun anaphora in Chinese texts*. Doctoral dissertation, City University of Hong Kong.

SIMPÓSIO 46

STRATEGIC IMPROVEMENT OF MACHINE TRANSLATED TEXTS BASED ON MANUAL EVALUATION

Francisco OLIVEIRA[280], Fai WONG[280], Sam CHAO[18]

ABSTRACT: The quality of translated documents generated by Machine Translation (MT) systems is always dependent on the domain and knowledge. Several measures have been proposed in the literature, which are mainly divided into Automatic and Manual evaluation metrics. In automatic evaluation, BLEU, NIST, Word Error Rate (WER), and METEOR are widely used in the community. However, they may not always reflect the actual errors found in the translated results. On the other hand, manual evaluation results are often based on the adequacy and fluency. In order to improve the translation quality strategically, this paper presents a manual evaluation of sentences generated by Portuguese Chinese Machine Aided Translation System (PCT) in terms of the following criteria: orthography, lexical selection, syntactic structure, concordance, verbal valence and related criteria. Solutions targeted for each issue are presented in details and these can improve the quality for similar cases in future translations generated by the system.

KEYWORDS: Machine Translation; Automatic and Manual Evaluation

1. Introduction

Translation task has triggered an enormous demand nowadays. As more documents have to be translated every day, the translation task becomes impractical without the help of Machine Translation (MT) systems. In the literature, different designs have been proposed. Rule based MT (Bennett et al., 1985) approach is based on a set of linguistic grammar rules for handling the translation. Example based MT (Brown, 1996; McTait, 2001) analyzes different pieces of bilingual examples stored in parallel corpora for generating the translation. Statistic based approaches (Brown et al., 1990; Lopez, 2008) relies on the probabilities estimated between the translation of words and the ordering of the sentences extracted from the corpora. However, each of these approaches has its strengths and weaknesses in the development of good MT systems.

In order to evaluate their translation quality, automatic and human evaluations are considered in this field. In Automatic Evaluations, translated sentences are compared to reference translations by measuring their overlapping approximations. The scoring function in BLEU (Papineni et al., 2002) evaluation metric is based on n-gram co-occurrence between the generated translation and the reference translations. NIST (Doddington, 2002) is based on BLEU but focuses on how informative a particular n-gram is. In other words, if an n-gram matched only occurs rarely, it receives more weight compared to another n-gram which often occurs. Word Error Rate (WER) (Sonja et al., 2000) is based on the Levenshtein distance, which measures the distance between the strings in terms of the number of edits required to change from one string into the other based on the number of operations, including insertion, deletion, and substitution. METEOR (Satanjeev et al., 2005) considers not only considers unigram precision and recall but also the score of the best pairing of the MT output with each reference by establishing their alignment relationships. Moreover, it also considers word inflection variations and synonyms for matching, and applies a word-ordering penalty instead of relying on higher *n*-gram matches.

On the other hand, in manual evaluations, generated translations are usually rated by users in terms of adequacy and fluency. In Automatic Language Processing Advisory Committee (ALPAC) (Carroll, 1966), human assessment is based on (1) intelligibility, which measures how "understandable" the sentence is, and (2) fidelity, which measures how much information the translated sentence is retained compared to the original. On the other hand, in Defense Advanced Research Projects Agency (DARPA) (White et al., 1994), it is based on adequacy, fluency, and informativeness. Adequacy rates how much information is transferred between the generated and reference translation. Fluency rates either the translation is fluent or not, and informativeness determines either enough information was conveyed in MT output to enable evaluators to answer questions on its content.

In Macau, a Portuguese-Chinese Machine Aided Translation System (PCT) (Oliveira et al., 2010) is used by the society in providing a translation workbench for professional translators to handle the daily translation work. It is a hybrid translation system which applies Translation Corresponding Tree (TCT) (Wong et al., 2006) as the Example based MT paradigm to represent the structure between the source and target for searching and matching the fragments between bilingual texts, and the application of Constraint Synchronous Grammar (Wong et al., 2006) as the Rule based paradigm to formulate the relationship between the languages simultaneously based on semantic constraints defined.

In order to improve the translation quality of this system strategically, this paper presents a manual evaluation, from the viewpoint of linguistics, based on the different criteria: orthography, lexical selection, syntactic structure, grammatical concordance, and verbal valence in the identification of errors committed by the system.

This paper is organized as follows. Section 2 introduces in details on each of the criteria considered, and their corresponding solutions to improve the translation quality. Section 3 shows some manual evaluation results and how much the system can improve after solving up the problems found. At last, a conclusion is followed.

2. Evaluation criteria

Typical evaluation strategies do not specify clearly what the type of error committed is. In most of the cases, they only reflect either the sentence is translated fluently and adequately or not, and the percentage of matches between the translation against the reference. In this section, different criteria are considered in better revealing the type of error in the generated sentence.

2.1 Orthography Errors and words which do not have any translation

Orthographic errors indicate either there are any translation results with wrong spellings. Although this rarely happens, since dictionaries are manually constructed, there is a chance to have some wrong spellings. On the other hand, for some new words or phrases which do not exist in our knowledge base, no doubt, they will not be translated.

2.2 Lexical Selection

Words are translated with different meanings according to the context. It is always considered as a difficult task for the machine to select the most appropriate text due to the inherent ambiguities of the languages.

2.3 Syntactic Structure

This type of error typically occurs when the knowledge base is out of domain. Since MT systems are highly dependent on the knowledge they apply, they can only generate a good syntactic structure only if proper rules are defined. In general, better syntactic analysis results can be generated when there is more knowledge, but it also increases the handling and analysis time.

18 Faculty of Science and Technology, University of Macau
Av. Padre Tomás Pereira, Taipa, Macau
{olifran, derekfw, lidiasc}@umac.mo

2.4 Grammatical Concord

Depending on the language itself, different types of grammatical concord have to be considered carefully in MT systems. Grammatical concord refers to the agreement of a number, person, and gender in the sentence context. As an example, in the Chinese sentence "一枝筆" (one pen), MT system do not have to consider the agreement relationships, but for the Portuguese sentence "uma caneta" (one pen), the word "um" must be linked with "a" to have the female gender in grammatical concord with the subject word "caneta" (pen).

2.5 Verbal Valence

By definition, verb valence indicates the number of arguments bounded by a verbal predicate. In order to establish a correct valence, MT systems may be required to consult the subject and the object in the context to generate a valid verb valence. Typically, restorations of verbs according to the subject and time have to be handled by MT systems. As an example, if the Chinese sentence "我做了功課" (I did the homework) is being translated to Portuguese, the verb "fazer" (to do) should be changed to first person, past tense by the system in the generation of "Eu fiz o trabalho de casa" (I did the homework). Furthermore, some verbs, according to the context, after translation, should be followed by some specific words to make the valence correct. If the verb "乘" (to take) is to be translated into Portuguese under the sentence "我乘自行車" (I take the bicycle), besides having a correct tense and number agreements, the sense associated to the subject should be human, and the object should be transportation tool. Moreover, in the Portuguese translation, the verb should follow the preposition "de" in the generation of "Eu ando de bicicleta" (I take the bicycle). Usually, when MT systems translate longer sentences, the difficulty in having a correct valence increases since they have to correctly identify long distance dependencies between the verbs, subjects, and objects in the context.

2.6 Solutions

Based on the evaluation criteria concluded, different solutions are applied to improve the PCT system. For handling orthographic errors and words which do not have translation, the easiest solution is to modify directly in the knowledge or add them back to the system. In order to ensure the correctness in the lexical selection, we may require to add or change senses in the knowledge base to guarantee the translation quality. If system encounters translations with wrong syntactic order or lexical selection, we will add proper CSG Part-of-Speech post-correction and Word Sense Disambiguation rules. Similarly, grammatical concord and valence issues can be handled in the same fashion.
Typically, for different languages, besides the mentioned errors, there could be other types of errors, and different solutions should be provided. As an example, for Chinese to Portuguese translation, if the segmentation of Chinese words in the sentence is done incorrectly, we have to also consider post-correction rules.

3. Evaluations

Experiments are conducted on a case study to evaluate our system in terms of the criteria discussed, and we tried to compare the system before and after the improvements are made. In Figure 1, it shows the types of errors committed in this case study generated by our system. Most of the errors are related with lexical selection (39%) and syntax structure (22%).

Table 1. Percentage of errors committed in the case study

Type of error	Percentage	Type of error	Percentage
Orthography	4%	Grammatical Concord	17%
Without any translation	9%	Syntax Structure	22%
Verbal Criteria	9%	Lexical Selection	39%

Based on the evaluation results, we tried to improve our system by taking different solutions: lexicon changes in the Knowledge base, addition and modification of CSG, Part-of-Speech, and Word Sense Disambiguation rules. At the end, we compared the translation quality before and after the changes are applied, as shown in Table 2.

Table 2. Automatic evaluation results before and after improvements are considered

	BLEU	NIST
Before	0.2476	5.1435
After	0.4832	6.9727

The translated result is compared with 13 different references. In terms of BLEU evaluation metric, we have around a 22% improvement after changes are applied into the PCT system. It is not easy to achieve a full score, i.e. 100% accuracy in the automatic evaluations in terms of translation quality. This is related with several issues. Firstly, since most of the evaluation metrics rely on *n*-gram co-occurrence precision, low scores can be obtained even if the translated sentence is correct but different wordings are used compared to reference translations. Secondly, long range dependencies are difficult to be handled at one stage by MT systems. Table 3 shows an example.

Table 3. Translation Example

Source sentence	綠茶, … , … , …保存期限也是最短的
Generated translation by PCT System	O chá verde ... , o **Ø** prazo de validade também é mais curto
Reference Translation	O chá verde ... , o ***seu*** prazo de validade é também mais curto

Since the word "綠茶" (green tea) appeared at the beginning of the source sentence, after several short sentences separated by comma, translators know that it is necessary to add the word "seu" (its) before "prazo de validade" (validity). However, for machine point of view, since there is no direct evidence in having "它的" (its) from the source text, the corresponding translation does not appear in the generated translation.

4. Conclusion

In this paper, a deeper evaluation is conducted to reflect the actual errors found in the translated results generated by the PCT System. From the viewpoint of linguistics, the following criteria are considered: orthographic errors, lexical selection, syntactic structure, concordance, and verbal valence. Based on these results, different solutions are proposed to improve the system strategically. Evaluations are conducted to show the percentage of improvement the system can obtain by comparing the translation results before and after the changes are applied.

Aknowledgements

This work was partially supported by the Research Committee of University of Macau under grant UL019B/09-Y3/EEE/LYP01/FST, and also supported by Science and Technology Development Fund of Macau under grant 057/2009/A2.

References

Bennett, W.S., Slocum, J.. 1985. *The LRC Machine Translation System*. Computational Linguistics 11(2-3), 111-121.

Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.. 1990. *A Statistical Approach to Machine Translation*. Computational Linguistics 16(2), 79-85.

Brown, R.D.. 1996. *Example-Based machine translation in the Pangloss system*. Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pp. 169--174. Copenhagen, Denmark.

Carroll, J.. 1966. *An experiment in evaluating the quality of translations*. Pierce, J. Language and Machines: Computers in Translation and Linguistics.

Doddington, G.. 2002. *Automatic evaluation of machine translation quality using n-gram co-occurrence statistics*. Proceedings of the Second International Conference on Human Language Technology Research, pp. 138-145, San Diego, California.

Fai Wong, Ming Chui Dong, and Dong Cheng Hu. 2006. *Machine Translation Based on Translation Corresponding Tree Structure*. Tsinghua Science and Technology, 11(1), pp. 25-31.

Fai Wong, Ming Chui Dong, and Dong Cheng Hu. 2006. *Machine Translation Using Constraint-Based Synchronous Grammar*. Tsinghua Science and Technology, 11(3), pp. 295-306.

Francisco Oliveira, Fai Wong, Sam Chao, Chi-Wai Tang. 2010. *Portuguese Chinese Machine Aided Translation System*. The Anthology of Selected Papers from FIT 6th Asian Translator's Forum, Macau.

Lopez, A.. 2008. *Statistical Machine Translation*. ACM Computing Surveys, Vol. 40, No.3, Article 8.

McTait, K.. 2001. *Translation Pattern Extraction and Recombination for Example-Based Machine Translation*. PhD Thesis, Centre for Computational Linguistics, Department of Language Engineering, UMIST.

Papineni, K., Roukos, S., Ward, T., Zhu, W.J.. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311--318, Philadelphia, Pennsylvania.

Satanjeev Banerjee and Alon Lavie. 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*. Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.

Sonja Niessen, Franz Josef Och, Gregor Leusch, Hermann Ney.. 2000. *An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research*. Proceedings of the 2nd International Conference on Language Resources and Evaluation, pp. 39-45, Athens, Greece.

White, J., O'Connell, T., O'Mara, F.. 1994. *The ARPA MT evaluation methodologies: evolution, lessons, and future approaches*. Proceedings of the 1994 Conference, Association for Machine Translation in the Americas, Columbia, Maryland.

TEORIA DA RELEVÂNCIA: UMA ANÁLISE DAS LACUNAS CULTURAIS NA TRADUÇÃO DE Viver

LI HUANG [19]

RESUMO: A Teoria da Relevância (Sperber e Wilson 1986/1995) fez uma grande contribuição para o campo linguístico. Trata-se de uma abordagem pragmático-cognitiva que toma por base a característica inerente à cognição humana. O seu Princípio de Relevância, considera que todo o acto de comunicação é ostensivo e comunica a presunção de sua relevância óptima. Gutt (1991/2000) aplica este princípio nos Estudos de Tradução, como um processo ostentivo-inferencial, cujo objectivo é atingir a relevância óptima. Devido à falta de conhecimento do universo do texto literário da língua-fonte pelos leitores da língua-alvo, é observado o fenómeno da lacuna cultural na tradução. Quanto à tradução das lacunas culturais, a selecção da estratégia é importante para atingir a relevância óptima de acordo com o critério de consistência do Princípio de Relevância, isso significa que a tradução deve oferecer aos leitores da língua-alvo efeitos contextuais suficientes sem impor a eles um maior esforço de processamento, para que não sobrecarregue o seu processamento cognitivo. Esta dissertação realizou uma análise das lacunas culturais na tradução do chinês para o português do romance *Viver* (Yu 1993/2008), de modo a observar se as estratégias adoptadas na tradução destas lacunas culturais chegaram à relevância óptima pelo critério da consistência do Princípio de Relevância.

PALAVRAS-CHAVE: efeito contextual; esforço de processamento; relevância óptima; lacuna cultural; tradução chinês-português

Teoria da Relevância

Numa conferência organizada pela Universidade de Harvard em 1967, o filósofo Grice proferiu uma palestra intitulada "*Logic and Conversation*", provocando um impacto para a área da linguística. "A ideia básica subjacente é que existe um hiato entre a construção linguística do enunciado pelo falante e a sua compreensão pelo(s) ouvinte(s). Esse hiato no processo interpretativo deveria ser preenchido não mais por decodificação e sim por inferências" (Silveira e Feltes 1999:21). Nesta perspectiva, Grice (1975) defende que existe um acordo entre falante e ouvinte, que o denominou de Princípio de Cooperação, o qual está relacionado a quatro categorias, para possibilitar uma comunicação bem-sucedida.

Segundo o filósofo, o Princípio de Cooperação e suas máximas são constituídas como o seguinte:

> *Princípio de Cooperação:*
> *Faça sua contribuição conversacional tal como é requerida no momento em que ocorre, pelo propósito ou direção do intercâmbio conversacional em que você está engajado.*[20]
> *Categorias e Máximas:*
> *A. Quantidade*
> *a) Faça a sua contribuição tão informativa quanto é requerido.*
> *b) Não faça a sua contribuição mais informativa do que é requerido.*
> *B. Qualidade*
> *a) Não diga aquilo que você acredita ser falso.*
> *b) Não diga aquilo para o qual você não dispõe de evidência adequada.*
> *C. Relação*
> *Seja Relevante.*
> *D. Maneira*
> a) *Evite obscuridade de expressões.*

19 BLCU, Instituto das Línguas Estrangeiras, Departamento de Português, Nº 15 da Rua Xueyuan, Distrito Haidian, 100083, Beijing, China, lhlhby622@yahoo.cn.

20 Traduzido por Silveira e Feltes (Op. Cit.): "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged."

b) *Evite ambiguidade.*
c) *Seja breve.*
d) *Seja ordenado*[21]
(Grice 1975: 45 apud Silveira e Feltes 1999: 22)

No Princípio de Cooperação, a concepção da relevância apresenta-se como uma das quatro máximas. Sperber e Wilson (1986/1995) desenvolveram a Teoria da Relevância (TR), a partir da Máxima da Relevância deste Princípio: "Uma suposição é relevante num contexto na condição de que essa tenha alguns efeitos contextuais naquele contexto"[22] (Sperber & Wilson 1986: 122).
Para elaborarem uma definição mais sistemática e uma explicação mais compreensível, elas introduziram as duas condições seguintes:

> *Condição 1: uma suposição é relevante no contexto à medida que houver um maior número de efeitos contextuais.*
> *Condição 2: uma suposição é relevante no contexto na medida em que o esforço para processá-la, nesse contexto, for pequeno*[23] (Sperber e Wilson 1986: 153)

De acordo com Sperber e Wilson (1986/1995), uma suposição que não tenha nenhum efeito contextual num determinado contexto não é relevante naquela situação comunicativa. Por outras palavras, ter algum efeito contextual dentro de um contexto é uma condição necessária para a relevância.
A partir do ponto de vista de Sperber e Wilson, a relevância é o factor mais importante na comunicação. Em relação à relevância, Sperber e Wilson consideram assim:

> *A Teoria da Relevância pode ser vista como uma tentativa de resolver em detalhe uma das afirmações centrais de Grice: a de que uma característica essencial da maior parte da comunicação humana, verbal e não-verbal, é a expressão e o reconhecimento de intenções* (Sperber e Wilson 2005: 220).

De acordo com a Teoria da Relevância, uma comunicação bem sucedida não ocorre apenas por seguir as máximas do Princípio da Cooperação ou por seguir outras regras comunicativas, mas sim através da busca pela relevância, que é uma característica básica da cognição humana. Romão considera:

> *A teoria da relevância é uma nova abordagem da pragmática que tenta dar resposta não só às questões filosóficas que se relacionam com a natureza da comunicação, mas também às questões psicológicas que dizem respeito ao modo como o processo da interpretação se desenrola na mente do ouvinte* (Romão 2008: 23).

21 Traduzido por Silveira e Feltes (Op. Cit.):
A. Quality: Try to make your contribution one that is true.
(a) Do not say what you believe to be false.
(b) Do not say that for which you lack adequate evidence.
B. Quantity
(a) Make your contribution as informative as is required for the current purposes of the exchange.
(b) Do not make your contribution more informative than is required.
C. Relation: Be relevant.
D. Manner: Be perspicuous.
(a) Avoid obscurity of expression.
(b) Avoid ambiguity.
(c) Be brief (avoid unnecessary prolixity).
(d) Be orderly.
22 An assumption is relevant in a context if and only if it has some contextual effect in that context.
23 Traduzido por Silveira e Feltes (1999, 44).
Extent condition 1: a phenomenon is relevant to an individual to the extent that the contextual effects achieved when it is optimally processed are large.
Extent condition 2: a phenomenon is relevant to an individual to the extent that the effort required to process it optimally is small.

Sperber e Wilson (1986/1995) elencam dois conceitos importantes para determinar a relevância: o efeito contextual e o esforço de processamento. Na comunicação, o contexto é construído conforme a situação. Conforme elas, o contexto é:

> *... um conjunto de premissas utilizadas para a interpretação de um enunciado... Um contexto não está limitado à informação sobre o ambiente físico imediato ou aos enunciados imediatamente precedentes, mas se relaciona também quanto às expectativas sobre o futuro, às hipóteses científicas ou às crenças religiosas, às memórias anedóticas, às suposições culturais gerais, às crenças sobre o estado mental do falante—tudo isso podem ter um papel na interpretação*[24] (Ibidem: 15-16).

"O contexto é definido como o conjunto de premissas—informações mentalmente representadas—que é utilizado para interpretar enunciados." (Silveira e Feltes 1999: 28) Um contexto é construído psicologicamente, e também é um subconjunto das suposições do ser humano sobre o mundo. O mundo não se refere ao mundo real e objectivo, mas sim, ao mundo construído com base nestas suposições. Sperber e Wilson ressaltam que o contexto selecionado para interpretar um enunciado é restringido pela organização da memória enciclopédica do indivíduo, pelas suas habilidades perceptuais e outras habilidades cognitivas, bem como pela atividade mental na qual está engajado naquele momento (Silveira e Feltes 1999: 47). Gutt ainda complementa:

> *O contexto não se refere a algumas partes do ambiente externo dos interlocutores da comunicação, quer proceder o texto quer seguir um enunciado, quer as circunstâncias situacionais quer os factores culturais, etc.; o contexto refere-se à parte das suas suposições sobre o mundo ou o ambiente cognitivo*[25] (Gutt 1991/2000: 26).

Gutt (1991/2000) refere que o ambiente cognitivo como as informações que as suposições oferecem aos destinatários e a disponibilidade do conhecimento dessas informações para o processo de interpretação. Estas informações podem ser percebidas no ambiente físico, recuperadas pela memória, etc. O contexto é importante na TR, através do qual se estabelece um ambiente cognitivo, construindo uma relevância entre os interlocutores. De acordo com Silveira e Feltes (1999: 47), "a seleção contextual é guiada pela busca da relevância no processamento da informação. Se os efeitos contextuais adequados forem alcançados com o mínimo de esforço justificável, então a informação terá sido optimamente processada."

1.1 Princípio de Relevância

Sperber e Wilson levantaram o Princípio de Relevância, que se chama o Princípio Comunicativo da Relevância: "Todo acto de comunicação ostensiva comunica a presunção de sua relevância óptima"[26] (1986: 158).

> *O corolário do Princípio é:*
> *a)* *ele se aplica a todas as formas de comunicação;*
> *b)* *os indivíduos cujo ambiente cognitivo o comunicador está tentando modificar são os destinatários do ato de comunicação;*
> *c)* *ele não garante que a comunicação, apesar de tudo seja sempre bem-sucedida.* (Silveira e Feltes 1999:52)
> d)

O Princípio de Relevância defende, por conseguinte, uma presunção de relevância óptima de um enunciado, as

24 The set of premises used in interpreting an utterance. A context in this sense is not limited to information about the immediate physical environment or the immediately preceding utterances: expectations about the future, scientific hypotheses or religious beliefs, anecdotal memories, general cultural assumptions, beliefs about the mental state of the speaker, may play a role in interpretation.
25 Context does not refer to some part of the external environment of the communication partners, be it the text preceding or following an utterance, situational circumstances, cultural factors, etc.; it rather refers to part of their assumptions about the world or cognitive environment.
26 Traduzido por Silveira e Feltes (1999: 52) do original: "Every act of ostensive communication communicates the presumption of its own optimal relevance." (Sperber e Wilson 1986: 158)

presunções são expressas assim:

> *(a) Um conjunto de suposições I, que o comunicador pretende tornar manifesto ao destinatário é relevante o suficiente para merecer que a audiência processe o estímulo ostensivo.*
> *(b) O estímulo ostensivo é o mais relevante que o comunicador poderia ter usado para comunicar I*[27] (Ibidem).
> *(c)*

A primeira presunção da relevância óptima, conforme a explicação de Goldnadel e Oliveira (2009:38), significa:

> *"O estímulo deve produzir os efeitos que compensem o esforço exigido para o seu processamento, e aos destinatários, este estímulo deve valer a pena ser processado. Também tem uma expectativa de que o seu resultado seja mais relevante possível com o estímulo."*

A presunção (b) estabelece uma exigência mais explícita, que é o mais relevante do estímulo com as suposições estabelecidas.

1.2. Princípio de Relevância e a tradução

He e Zhang (2001: 289) explicam o modelo de processo da tradução da TR através do seguinte diagrama:

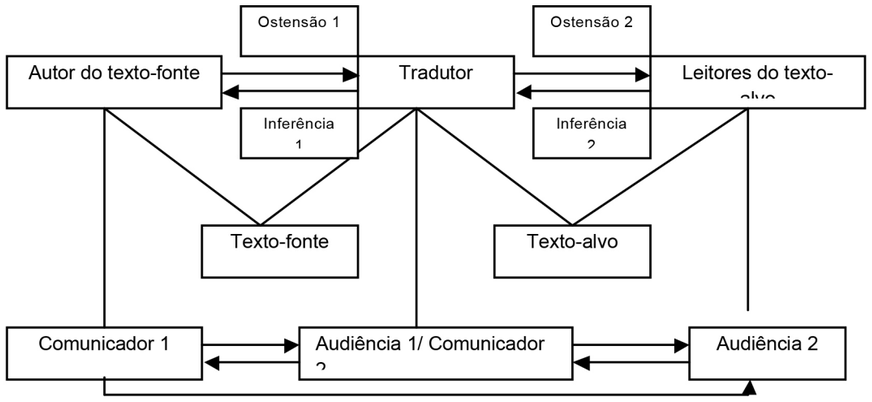


Diagrama 1: Modelo de processo de tradução da TR proposto por He e Zhang (2001)

Conforme He e Zhang (2001), o acto da tradução consiste em dois processos ostensivo-inferenciais. Três comunicadores são envolvidos neste processo: o autor do texto-fonte, o tradutor e os leitores-alvo. No primeiro processo ostensivo-inferencial, o tradutor infere a intenção do autor através do contexto, seguindo provavelmente o Princípio de Relevância. Depois do primeiro processo, o tradutor torna-se um comunicador no segundo processo que manifesta as intenções do autor aos leitores-alvo de um modo ostensivo e de acordo com a sua inferência. O tradutor, a partir de seu aparato cognitivo, escreve uma versão segundo o seu julgamento quanto as habilidades perceptuais e as expectativas dos

27 Traduzido por Silveira e Feltes (1999: 51) do original: "
Presumption of optimal relevance
The set of assumptions I which the communicator intends to make manifest to the addressee is relevant enough to make it worth the addressee´s while to process the ostensive stimulus.
The ostensive stimulus is the most relevant one the communicator could have used to communicate." (Sperber e Wilson 1986: 158)

leitores-alvo, para que esses possam entender as intenções do texto-alvo com o mínimo esforço de processamento, a fim de alcançar uma relevância óptima. Os leitores-alvo inferem essas intenções através das informações ostentas pelo tradutor.
Na tradução, entre as várias interpretações possíveis duma palavra ou duma frase, o tradutor selecciona apenas uma, que julga ser a mais adequada ao seu interlocutor. Conforme Silveira e Feltes (1999: 53), "é esse critério – o da consistência com o Princípio de Relevância – que garante a seleção de uma única interpertação para o enunciado entre as várias interpretações possíveis".

1.3. Lacuna cultural

Conforme Newmark (1981/2001: 94), a cultura é "como o modo de vida, e as suas manifestações são peculiares numa dada comunidade que usa uma linguagem particular como o seu meio de expressão"[28]. A cultura reflecte-se nos diversos aspectos da vida, Newmark (2001:103) categoriza a cultura nos seguintes aspectos:

(1) Ecologia: animais, plantas, montanhas, planícies, clima, etc.;
(2) Cultura material: alimentação, vestuário, habitação, transporte e comunicação;
(3) Cultura social: trabalho e lazer;
(4) Organizações, costumes e ideias: a nível político, social, jurídico, religioso e artístico.
(5) Gestos e hábitos.

1.3.1. . Definição da lacuna cultural

Dagut (1978:45) define o termo lacuna como " a não-existência de uma equivalência de um termo duma língua em outra"[29].
Ele identifica quatro tipos de lacunas e propõe estratégias para supri-las:

a) ambiental: emerge a partir da não-traduzibilidade dos fenómenos naturais. O modo mais eficaz para resolver este tipo de lacuna, segundo o estudioso, é através da transliteração.
b) cultural: é dividida em religião e secularização, a estratégia de transliteração é acompanhada com uma nota de rodapé.
c) lexical: embora não tenha um termo equivalente, a conotação que o termo-fonte carrega pode existir na experiência dos leitores da língua-alvo. Para este tipo de lacuna, três técnicas de tradução são sugeridas: 1) selecciona-se um termo na língua-alvo que carrega uma parte da conotação do termo na língua-fonte; 2) faz-se a paráfrase de algumas características do termo na língua-fonte; 3) faz-se a omissão.
d) sintática: "é causada pelas assimetrias estruturais entre uma dada língua-fonte e uma dada língua-alvo[30], o autor sugere o ajuste sintáctico para chegar a uma equivalência. (Dagut 1978: 89).
e)
f)

Conforme Wang (1997: 55), a lacuna cultural é "uma falta de conhecimento do ambiente cultural relevante e compartilhado entre o autor e os seus leitores pretendidos"[31]. Também aponta que "a lacuna cultural é um fenómeno da comunicação de cultura-específica, que resulta do movimento de uma cultura particular"[32] (ibidem).
A lacuna na tradução é, conforme Gutt (1991/2000), "devido à diferença do ambiente cognitivo; pois, aos leitores da língua-alvo falta o conceito ou informação suficiente associado à concepção da língua-fonte"[33]. Também considera que

28 The way of life and its manifestations that are peculiar to a community that uses a particular language as its mean of expression.
29 Non-existence in one language of a one-word equivalent for a designatory term found in another.
30 These are caused by "structural asymmetries between SL and TL".
31 Absence of relevant cultural background knowledge shared by the author and his/her intended reader.
32 Cultural default is a culture-specific communication phenomenon and it is a result of the movement of a particular culture.
33 Due to differences in cognitive environment, the receptor language audience lacks information associated with a concept in the

a lacuna é inevitável na tradução (ibidem: 120). Para que o conceito do texto-fonte seja transmitida adequadamente aos leitores-alvo, a selecção das estratégias da tradução é importante.

1.3.2. Domesticação e estrangeirização

Nos Estudos da Tradução, a domesticação e a estrangeirização são dois lados opostos quanto à estratégia. Conforme Venuti (2004: 120), "a tradução imita os valores linguísticos e literários de um texto estrangeiro, mas a imitação é moldada numa língua diferente que se relaciona a uma tradição cultural diferente"[34].

A domesticação é uma técnica da tradução que visa eliminar, substituir ou simplificar os elementos que possam prejudicar o entendimento dos leitores-alvo, de modo a facilitar a leitura.

Os leitores-alvo, sendo o foco da domesticação, através da tradução, vão ter uma leitura fluente no seu contexto. Venuti (2004) considera que a linguagem do texto traduzido deve ser natural e fluente, correspondendo ao hábito dos leitores-alvo.

Conforme Campos (2009: 70), a estrangeirização "privilegia o contexto fonte, ou seja, o leitor é levado até o texto original pela manutenção das características linguístico-culturais do texto-fonte." A estrangeirização possibilita a oportunidade aos leitores-alvo a adquirir conhecimentos novos duma outra cultura. "Assim, quanto mais se evidenciar a estrangeiridade do texto, maior a oportunidade de se desenvolver um público-leitor mais aberto às diferenças linguísticas e culturais." (Abreu 2010:154)

2. Análise das lacunas culturais na tradução de Viver

Yu Hua, autor desse romance, nasceu em 1960 em Hangzhou, capital da província de Zhejiang. Foi o primeiro escritor chinês a receber o Prémio da Fundação James Joyce em 2002. *Viver*, considerado um dos romances chineses mais influentes da década de noventa, ganhou o prémio italiano Grinzane Cavour em 1998 e foi publicado em mais de treze países. Em 1994, o realizado chinês Zhang Yimou adaptou-o para um filme sob o nome de *Tempos de Viver*, que foi aclamado pela crítica e pelo público.

O romance passa-se no início dos anos de quarenta, onde a família *Xu*, que residia numa aldeia do interior da China, gozava duma vida confortável e próspera. Os seus antepassados acumularam riqueza o suficiente para o conforto da família, podendo sustentar às gerações posteriores. Ao fim da Segunda Guerra Mundial e da ocupação japonesa, quando as tropas nacionalistas tratavam de recuperação das terras usurpadas pelo invasor, Fugui, filho único da família *Xu*, perdeu tudo por causa do seu gosto pelo jogo e pelas mulheres, deixando a família numa enorme dívida. A vida caiu numa pobreza absoluta. Para sustentar a família, Fugui conseguiu obter uma parcela de terra e começou a vida nova. *Viver* é um romance que conta a trajectória da China contemporânea através da descrição épica da família *Xu*.
Márcia Schmaltz, tradutora deste romance, é natural do Brasil tendo trabalhado como tradutora e intérprete em português e em chinês na área comercial. É actualmente professora de língua portuguesa e tradução na Universidade de Macau. Ainda criança, morou por seis anos em Taiwan, em consequência do casamento da mãe com um chinês. A obra foi traduzida directmente do chinês para o português tendo sido publicada no Brasil em 2008. A experiência como tradutora e intérprete, a vivência em Pequim e o contacto constante com a China oferecem-lhe uma ampla visão que lhe ajuda muito na sua carreira da tradução.

original.
34 Traduzido por Abreu (2010)

2.1 Técnicas tomadas na tradução das lacunas culturais em Viver

No romance *Viver* (Yu 2008), a tradutora empregou tanto a domesticação como a estrangeirização perante diferentes tipos de lacunas culturais, que apresentamos nas subsecções seguintes.

2.1.1 Domesticação

A domesticação pretende aproximar o autor do texto original dos leitores da língua-alvo. Venuti (2004) afirma que esta técnica tem que ser fluente e clara, de modo a diminuir o estranhamento do texto-fonte que poderá causar aos leitores-alvo, conforme explanado na secção 5 do capítulo 2. A domesticação ainda pode empregar outras medidas tais como tradução por sentido, substituição, paráfrase, que são abordadas a seguir.

2.1.1.1 Tradução por sentido

TF: ，，，。(Yu 1993: 138)

TA: Eu, por minha vez, estava assustado e não conseguia dormir. Jiazhen parecia bem melhor, mas eu temia que fosse apenas uma recuperação aparente, o último brilho do sol antes do poente. (Yu 2008: 149)

"" referente ao fenómeno natural, remete a imagem do brilho temporário do céu, devido ao reflexo dos raios solares às nuvens, e, depois, cai rapidamente a escuridão. Esta expressão é metafórica, descrevendo o estado de ânimo que uma pessoa poderá ter antes de morrer.

Jiazhen estava doente, e não tinha força para ficar de pé. Conforme o diagnóstico do médico, ela iria morrer em breve. Contudo, um dia melhorou e conseguiu se levantar da cama e até fazer alguns trabalhos domésticos. Fugui ficou surpreso com a sua melhora repentina e o súbito falecimento, por isso utilizou este dito.
Conforme as três técnicas adoptadas por Dagut (1978), a tradutora fez uma interpretação apropriada e conhecida pelos leitores brasileiros, "o último brilho do sol antes do poente", através da qual os leitores-alvo, inferem a conotação de sentido com menos esforço de processamento, conforme o contexto em que se enquadra a expressão.

2.1.1.2 Substituição

A substituição, segundo Beekman e John (1974: 201), refere-se à utilização do termo existente na cultura-alvo para descrever o termo no texto-fonte, estes dois termos têm a mesma função.
TF: ，，。(Yu 1993: 11)
TA: O rosto de meu sogro ficava vermelho como um pimentão. Eu me afastava rindo. (Yu 2008: 16)

Figura 1: Imagem de ""[35]

35 Disponível em http://image.baidu.com/i?ct=503316480&z=&tn=baiduimagedetail&word=%CB%C9%BB%A8%B5%B0&in=10724&cl=2&lm=-1&pn=0&rn=1&di=36535556490&ln=2000&fr=&fmq=&ic=0&s=0&se=1&sme=0&tab=&width=&height=&face=0&is=&istype=2#pn0&-1, acessado em 02/07/2011.

Na base da classificação de cultura por Newmark (2001), o termo “” pertence à cultura material que é alimentação. Para que a tradução de “” seja aceitável na língua-alvo, a tradutora recorreu à substituição. Nesta frase, “” foi traduzido por “pimentão”.

Segundo Huang (2008), para traduzirmos este tipo de lacuna cultural, podemos empregar a técnica de transliteração com uma nota de rodapé para apresentar mais conhecimento sobre “”. Neste caso, a tradutora fez uma substituição. Como traduz adequadamente com esta técnica?

Conforme a narrativa, o Fugui é viciado em jogo e tinha recaída por mulheres. Mesmo ele estando casado com a Jiazhen, sempre ia para o prostíbulo. Além disso, sempre que levava uma outra mulher à loja do seu sogro, cumprimentava-o, fazendo com que este ficasse zangado e envergonhado.

Neste caso, o texto-fonte utilizou o recurso de *símile* para descrever a expressão do rosto do sogro do Fugui. A cor interna de “” é verde escura. Na cultura chinesa, é às vezes empregar esta expressão para denotar a expressão duma pessoa zangada, e em alguns casos, usa-se também diretamente a cor verde para expressar esta situação, como por exemplo “” (literalmente rosto verde como o ferro)[36]. O termo chinês indica a cor do rosto do sogro de Fugui depois de ver que Fugui levou uma outra mulher à sua loja.

Segundo *Dicionário Contemporâneo da Língua Portuguesa Caldas Aulete*, o termo pimentão é: “Diz-se de alguém que está com o rosto vermelho (por queimadura de sol, p.ex.) ou muito corado, ruborizado (de excitação, vergonha etc.)”. A expressão “vermelho como pimentão” descreve o rosto do sogro de Fugui, que “estava zangado e envergonhado com as atitudes do genro, é uma expressão da cultura brasileira”[37].

Dessa forma os dois termos manifestam o aspecto figurativo semelhante, que ambos descrevem a situação zangada do sogro de Fugui, conhecido como *símile*, relacionado com o contexto. A tradução oferece o efeito adequado aos leitores da língua-alvo que podem inferir a conotação deste termo relativo a esta frase sem desprender esforços desnecessários.

2.1.1.3 Paráfrase

Dryden (1989: 8) define a paráfrase como “a tradução com latitude, onde o tradutor toma em consideração o autor para que não perça o sentido original, contudo, a sua tradução não é tão estritamente conforme o original”[38]. Ele privilegia a paráfrase na tradução e argumenta que este método de tradução pode “ transferir o espírito do autor”[39] (ibidem: 11).
TF: ， ， ， ， ， 。(Yu 1993: 8)

TA: Pensando bem, eles estavam certos. No início, eu não aceitava. Eu pensava: tenho dinheiro, sou filho único, se eu bater as botas, lá se vai a família Xu. (Yu 2008: 13)

Na base da classificação culturais de Newmark (2001), “” é cultura social. Dentro as três técnicas propostas por Dagut (1978), “ 1) selecciona-se um termo na língua-alvo que carrega uma parte da conotação do termo na língua-fonte; 2) faz-se a paráfrase de algumas características do termo na língua-fonte; 3) faz-se a omissão.”, a tradutora empregou a segunda técnica, que é paráfrase.

“” refere-se ao lume dos incensos. Conforme o romance, o Fugui é “” da família Xu, isso significa que ele é o filho único da família Xu, que pertence ao costume (Newmark 2001) da China. Se o incenso for apagado (), a continuidade da linhagem familiar será interrompida. No critério da consistência com o Princípio de Relevância, a tradutora fez uma paráfrase e traduziu por “sou filho único”, seguida pela respectiva frase explicativa: “se eu bater as botas, lá se vai a família Xu”.

No *Dicionário Informal*, a expressão “bater as botas” possui “origem à primeira invasão holandesa, ocorrida em Salvador, em 1624, os negros comportaram-se bravamente diante do invasor. Não estavam acostumados com os armamentos que lhes foram dados, e constantemente tropeçavam nas próprias botas, virando um alvo fácil para os holandeses. Então os outros negros costumavam dizer que a pessoa havia batido as botas. Daí nasceu esta expressão, que significa morrer”.[40]
A tradução ostensiva manifesta o papel importante do Fugui na família *Xu*. Através da adaptação da expressão do TF na tradução, pretende-se oferecer os mesmos efeitos do TF aos leitores da língua-alvo:

a) Fugui é o filho único da família Xu.
b) Se ele bater as botas, interrompe-se a continuação da linhagem da família Xu.
c) Fugui exerce um papel muito importante na família Xu, conforme a) e b).
d)

Os leitores da língua-alvo podem perceber a tradução, porém, talvez não saibam esta cultura na China, que é apenas o filho consegue exercer a função da continuação da linhagem da família. Na tradução desta lacuna cultural, às vezes, não pode atingir a uma relevância óptima devido a esta diferença cultural profunda. Isso não depende da capacidade da tradutora.

2.1.2 Estrangeirização

A técnica de estrangeirização faz com que os leitores da língua-alvo aproximem-se do autor do texto original. Venuti (2004) explica que a estratégia de estrangeirização conserva os elementos exóticos do texto-fonte quando o tradutor faz a tradução, conforme já explanado na seção 1.3.2. Na tradução deste romance, há dois tipos de estrangeirização: a transliteração sem nota de rodapé e a com uma nota de rodapé.

2.1.2.1 Transliteração sem nota de rodapé

Conforme Huang (2008: 29), “há cada vez mais lacunas culturais podem ser transmitidas com suficiente intercâmbio cultural”[41], por isso, “algumas lacunas culturais podem ser traduzidas literalmente na língua-alvo sem nenhuma explicação”[42].

TF: ， ， ， ， 。(Yu 1993: 19)

TA: A Jiazhen daqueles tempos era muito bonita: o cabelo penteado rente atrás da orelha, com as pregas do *qipao* dançando na cintura enquanto ela caminhava. Naquele momento, pensei: quero que ela seja minha mulher. (Yu 2008: 25)

Figura 2: imagem de “”[43]

36 Disponível em http://zhidao.baidu.com/question/180397336.html?fr=qrl&cid=978&index=1&fr2=query, acessado em 07/07/2011.
37 Disponível em http://fanfiction.com.br/historia/53179/Dois_Mundos_Duas_Vidas/capitulo/32, acessado em 07/07/2011.
38 Translation with latitude, where the autor is kept in view with by the translator, so as never to be lost, but his words are not so strictly followede as his sense.
39 The spirit of an autor may be transferred.
40 Disponível em: http://www.dicionarioinformal.com.br/definicao.php?palavra=bater+as+botas, acessado em: 05/07/2011.
41 More and more cultural defaults would become transplantable with sufficient cultural exchange.
42 Some cultural default can be translated literally into target language text without any explanation.
43 Informação disponível em

SIMPÓSIO 46

Conforme *O grande dicionário da língua chinesa [辞海]* (1999: 2328), indica que “” é um vestuário feminino típico da China, que remonta ao vestuário tradicional da etnia *Manchu*, a figura de “” é mostrada na Figura 4. A tradutora fez uma transliteração, introduzindo esta concepção aos leitores de língua-alvo. Neste caso, a tradução fez uma ostensão de que o termo na língua-fonte não existe na cultura da língua-alvo. Os leitores-alvo têm de fazer inferências através desta tradução:

a) “” é um vestuário que modela o corpo, conforme as palavras “prega”, “cintura”.
b) “” é um vestuário feminino, conforme a frase “ dançando na cintura enquanto ela caminhava”.

Baseando nas categorias culturais de Newmark (2001), “” é um vestuário que pertence à cultura material, Dagut (1978) e Huang (2008) propõem a estratégia de transliteração acompanhada com uma nota de rodapé.

Mas, neste caso, a tradução literal “*qipao*” não é suficiente para os leitores da língua-alvo capturarem a imagem referenciada pelo item lexical. Embora a tradução seja orientada pelo critério da consistência com o Princípio de Relevância, esta tradução não permitiu a procura duma relevância óptima do leitor, por faltar informações contextuais suficientes para a inferência do LA. Se adicionarmos uma nota de rodapé, mais efeitos poderão ser inferidos pelos leitores-alvo.

2.1.2.2 Transliteração com nota de rodapé

Huang (2008: 30) refere que “alguns textos com lacunas culturais são muito específicos e é impossível ser aceitos directamente dos leitores da língua-alvo, neste caso, esta técnica é effectiva”[44]. Na tradução deste romance, a tradutora empregou a transliteração mais uma nota de rodapé em alguns casos.

TF: ， ， ， 。(Yu 1993: 9)

TA: Jiazhen estava um tanto deformado em razão da gravidez de seis meses de Youqing; ela caminhava como se tivesse um *mantou** no meio das pernas, que a fazia andar que nem uma pata com as pernas abertas. (Yu 2008: 15)

*Pão chinês feito à base de farinha de trigo, fermento e água e assado no vapor. (N.T.)

Figura 3: Imagem de “”[45]

“”, um dos alimentos populares na China, é um tipo de pão mas é feito a vapor em vez de forno. Os leitores da língua-alvo que não têm o conhecimento sobre o “”, espécie de pão, vão deparar-se com uma lacuna cultural aqui. Por isso, a estratégia de anotação através duma nota de rodapé fica bem neste contexto, para uma melhor compreensão. Através da anotação, os leitores-alvo também podem conhecer mais sobre a cultura chinesa.

Na base de Newmark (2001), “” é a alimentação que pertence à cultura material. Para este tipo de lacuna, Dagut (1978) sugere a estratégia de transliteração mais uma nota de rodapé para dar mais informações aos leitores da língua-alvo. A tradutora empregou esta técnica neste exemplo. Nesse romance, além de “”, outros termos da alimentação típica chinesa foram citados, como “”, “”, “” e “”. “” que são conhecidos é como “*tofu*”, queijo de soja correspondente à cultura dos leitores-alvo. “” ou simplesmente , um tipo de massa em forma de disco, feita de farinha, água e ovos e assada na chapa. “” é uma espécie de salgadinho chinês, recheado com carne e/ou verduras. “” é um tipo de licor chinês, destilado principalmente do sorgo, ao qual podem ser adicionados outros grãos. Com estas anotações, os leitores da língua-alvo entendem o sentido do texto-fonte com menos esforço de processamento.

3. Conclusão

Este trabalho tem como objectivo observar quais foram as estratégias adoptadas na tradução para tentar chegar à relevância óptima pelo critério da consistência pelo Princípio da Relevância e avaliar se a técnica empregada foi ostensiva o bastante, sem exigir maior esforço de processamento do leitor do TA.

Através da análise das traduções das lacunas culturais do romance *Viver*, chegamos à seguinte conclusão: A TR é uma teoria que se mostrou adequada para a explanação do processo de tradução de *Viver*. A análise foi aplicada conforme dois factores: o efeito contextual oferecido pela tradução e o esforço de processamento imposto pelos leitores da língua-alvo. Se estes puderem capturar efeitos contextuais adequados com menos esforço de processamento, estabelece uma relevância óptima conforme o Princípio de Relevância. Às lacunas culturais emergentes do romance, com distintas técnicas apropriadas, a relevância óptima não pode ser atingida cem por cento, seguindo as análises mostradas na secção 2.2. O sucesso da tradução das lacunas culturais não é apenas restringido pela capacidade do tradutor, mas também pelas diferenças linguística e cultural, especialmente as diferenças entre os dois países ou regiões com longa distância. Cada técnica tem a sua vantagem e disvantagem. Em alguns casos, a tradução não foi ostensiva o suficiente. A técnica de domesticação pode facilitar a compreensão dos leitores da língua-alvo, porém, perde às vezes a imagem figurativa carregada do termo do texto-fonte. A transliteração sem nota de rodapé pode apresentar novo conhecimento da cultura do texto-fonte, contudo, não oferece informações suficientes quanto a este termo, enquanto que a transliteração com nota de rodapé, embora ofereça as informações necessárias, irá influenciar o fluxo da leitura dos leitores.

Referências bibliográficas

Abreu, A. L. S.V. (2010) “Pollyanna: domesticação e estrangeirização na tradução de Monteiro Lobato”, in *Caderno do CNLF*, V.XIV, n.2, t.2, 1543-1554. Disponível em: http://www.filologia.org.br/xiv_cnlf/tomo_2/1543-1554.pdf, acessado em 3 de Julho de 2011.

Aulete, F. J. C. V., Santos, A. L. (2009) *Dicionário contemporâneo da língua portuguesa Caldas Aulete*. Edição brasileira original: Hamílcar de Garcia.

Araújo, T. X. L. (2005) *A tradução para o português brasileiro da ironia veiculada na obra Gulliver's Travels, de Jonathan Swift: uma análise à luz da Teoria da Relevância*, MA Dissertação, Manchester: St. Jerome Publishing.

Beekman, J. e John, C. (1974) *Translating the Word of God*, Grand Rapids, Michigan: Zondervan.

Campos, C. (2009) O pensamento e a prática de Monteiro Lobato como tradutor, in *IPOTESI—Revista de Estudos*

http://baike.baidu.com/image/3bb224879b64e264c65cc3a0, acessada em 15/05/2011.

44 Many texts with cultural defaults are so culture-specific that it is almost impossible for them to be accepted directly by the target reader. In this case, literal translation with notes should be an effective way to deal with the problem.

45 Informação disponível em: http://image.baidu.com/i?ct=503316480&z=&tn=baiduimagedetail&word=%C2%F8%CD%B7&in=14889&cl=2&lm=-1&pn=618&rn=1&di=8561165670&ln=2000&fr=&fmq=&ic=&s=&se=&sme=0&tab=&width=&height=&face=&is=&istype=#pn618&-1&di8561165670&objURLhttp%3A%2F%2Fimg5.poco.cn%2Fmypoco%2Fmyphoto%2F20080727%2F00%2F43887288200807270040091358837117215_002_640.jpg&fromURLhttp%3A%2F%2Fnj.food.poco.cn%2Ffoodiarydetail.php%3Fid%3D2413005%26stat_request_channel%3D3219320902&W450&H600, acessado em 25/07/2011.

Literários, V.13, n.1, 67-79. Disponível em http://www.revistaipotesi.ufjf.br/ipotes21.html. Acessado em: 3 de Julho de 2011.

Dagut, M. (1978) *Hebrew- English Translation: A Linguistic Analysis of Some Semantic Problems*. Haifa: University of Haifa.

Dryden, J. (1989) "Metaphrase, Paraphrase and Imitation". In Andrew Chesterman (Ed.), *Reading in Translation Theory [C]*, Helsinki: Finn Lectura, 7-12.

Bassnett S. (2005) *Estudos de tradução*. Traduzido por Sônia Terezinha Gehring, Letícia Vasconcellos Abreu e Paula Azambuja Rossato Antinolfi. Porto Alegre: Editora da UFRGS.

Godoi, E. e Santos e Sebestião, L. (2010) "Cognição e relevância: uma análise pragmática da interpretação inferencial de enunciados". In *Eletras*. V.20, n.20, 72-83.

Goldnadel, M. e Oliveira e R. de C. (2009) "Contribuição da teoria da relevância para a prática de interpretação de textos: uma ilustração por meio de textos de humor", in *Linguagem & Ensino, Pelotas*. V.12, n.1, 33-48.

Gonçalves, J.L.V.R. (2005) "Desenvolvimento da pragmática e a teoria da relevância aplicada à tradução", *Revista Linguagem em (Dis)curso*, V.5, n. especial, 129-150. Disponível em: http://www3.unisul.br/paginas/ensino/pos/linguagem/0403/00.htm, acessado em 15 de Maio de 2011.

Gutt, E.A. (2000) *Translation and Relevance: Cognition and Context*, Manchester & Boston: St. Jerome Publishing.
Hatim, B. e Mason, I (1990) *Discourse and the Translator*. London: Longman.
Ibarretxe-Antuñano, I. (2008) "Vision metaphors for the intellect: Are they really cross-linguistic?", in *Atlantis*, 30(1): 15-33.

Mahdavi, Z. A. (2006) "Relevance Theory and Explicitation Strategy in Translation", in *Translation Studies*. V.4, n14, 61-71.

Newmark, P. (2001) *A Textbook of Translation*, 上海：上海外与教育出版社 [Shanghai: Shanghai Foreign Language Education Press].

Nida, E.A. e Charles, R.T. (1969/1982) *The Theory and Practice of Translation*, Leiden: E.J.Brill.

Profissões Tradicionais Chinesas (2008). Macau: Instituto Português do Oriente e Instituto Politécnico de Macau.

Romão, S.C.G. (2008) *Do desafio do humor à sedução do processamento do texto humorístico à luz da teoria da relevância*, Tese de doutorado, Manchester: St. Jerome Publishing.

Silveira, J. R. C. da e Feltes, H.P. de M. (1999) *Pragmática e Cognição: a Textualidade pela Relevância e Outros Ensaios*, Porto Alegre: EDIPUCRS.

Sperber D. e Wilson D. (1986/1995) *Relevance: Communication and Cognition*, Oxford: Blackwell.

_____ (2005) "Posfácio da edição de 1995 de Relevância: comunicação e cognição" (tradução de Fábio José Rauen e Jane Rita Caetano da Silveira). In *Revista Linguagem em (Dis)curso*, V.5, n. especial, 171-220. Disponível em: http://www3.unisul.br/paginas/ensino/pos/linguagem/0403/00.htm, acessado em 10 de Maio de 2011.
_____ (2005) "Teoria da Relevância" (tradução de Fábio José Rauen e Jane Rita Caetano da Silveira), in *Revista Linguagem em (Dis)curso*, V.5, n. especial, 221-263. Disponível em: http://www3.unisul.br/paginas/ensino/pos/linguagem/0403/00.htm, acessado em: 10 de Maio de 2011.

Venuti, L. (2004) *The translator's invisibility*. 上海：上海外与教育出版社 [Shanghai: Editora da Educação da Língua Estrangeira de Shanghai].

Yu, H. (2008) *Viver* (tradução de Márcia Schmaltz). São Paulo: Companhia das Letras.

Cai, R.S. e Wang, Y. (蔡荣寿、王圆) (2007) "英汉俗语互译中的文化缺省及其翻译" ["As Lacunas Culturais na Tradução entre o Inglês e o Chinês"]. 《浙江传媒学院学报》[*Jornal do Instituto da Comunicação de Zhejiang*] V.3, 57-60

Cihai (辞海) (2000) [*Grande Dicionário da Língua Chinesa*]. 上海：上海辞书出版社 [Shanghai: Editora de CISHU de Shanghai].

Ciyuan (辞源) (1988/1997) [*Dicionário Etimológico Chinês*]. 上海：上海辞书出版社 [Shanghai: Editora de Cishu de Shanghai].

He, Z.R. e Zhang, X.H. (何自然、张新红) (2001) "语用翻译：语用学理论在翻译中的应用" ["Tradução Pragmática: Aplicação da Pragmática na Tradução"] 《现代外语》[*Língua Estrangeira Contemporânea*] V. 3, 285-293.

He, R. (何蓉) (2009) 关联理论关照下的《红楼梦》文化缺省翻译比较研究 [*Um Estudo Comparativo sobre a Tradução das Lacunas Culturais no Pavilhão Vermelho, à luz da Teoria da Relevância*]. Dissertação de Mestrado não publicada, Universidade de Guangxi.

Huang. H. (黄华) (2008) "从关联理论看文化缺省的翻译" 。["a Tradução das Lacunas culturais à luz da Teoria da Relevância"]. Dissertação de Mestrado não publicada, Universidade Normal de Guangxi.

Lin, K.N. (林克难) (1994) "关联理论翻译简介" [Introdução da Teoria da Relevância da Tradução]. 《中国翻译》[*Tradução na China*] V.4, 6-9.

Wang, D. F. (王东风) (1997) "文化缺省与翻译中的连贯重构" [Lacuna cultural e a Reconstrução da Coerência na Tradução]， In:《外国语》 [*Língua Estrangeira*]，, V.6, 55-60.

_____ (2000) 《文化缺省与翻译补偿》[*Lacuna Cultural e a Compensação da Tradução*] 北京：中国对外翻译出版公司 [Beijing: Editora da Tradução ao estrangeiro da China].

Yu, H. (余华) (1993) 《活着》 [*Viver*], 上海：山海文艺出版社 [Shanghai: Editora da Literatura e da Arte].

Zhang, X.H. (张晓红) (2008) "从关联理论看语用翻译" ["Tradução Pragmática à luz da Teoria da Relevância"] 《东北大学学报》[*Jornal da Universidade Nordeste*]，V.10-3, 269-273

Zhongguo xiehouyu dacidian (中国歇后语大辞典) (2011) [*Grande Dicionário do Dito Chinês*]. 上海：上海辞书出版社 [Shanghai: Editora de *Cishu* de Shanghai].