Fine-Tuning in Brazilian Portuguese-English Statistical Transfer Machine Translation: Verbal Tenses

Lucia Silva
University of São Paulo
Av. Prof. Luciano Gualberto, 403
Cidade Universitária, São Paulo, BR
helena.rozario@usp.br

Abstract

This paper describes an experiment designed to evaluate the development of a Statistical Transfer-based Brazilian Portuguese to English Machine Translation system. We compare the performance of the system with the inclusion of new syntactic written rules concerning verbal tense between the Brazilian Portuguese and English languages. Results indicate that the system performance improved compared with an initial version of the system. However significant adjustments remain to be done.

1 Introduction

Recently, Statistical Machine Translation systems have received much attention because they are fully automated and have shown significant improvements over other types of approaches. Experiments with string-to-string and syntax-based systems have shown better results when linguistic features are added to Machine Translation systems (Chiang et al., 2009). The Statistical Transfer (Stat-XFER) approach presented in this paper was designed as a Statistical approach with a grammar module, which encodes syntactic transfer rules, i.e., rules which encode constituent structures from the source language to the target language structure.

Verbal tenses vary among natural languages. Each language has its typical verbal form, and they share mood, voice, aspect and person qualities (Comrie, 1993). Some languages, such as Portuguese and English do not share the same properties and the number of verbal tenses may present divergences. Our goal is to test the development of the Statistical Transfer-based system under the application of syntactic transfer rules of verbal tenses involving this pair of languages.

2 The Statistical Transfer-based system

The hybrid Stat-XFER system uses a transfer-based method and statistical paradigm for the translation process and it is composed of the following main components: the Bilingual Phrasal Lexicon, Morphology, Transfer Engine and the Decoder. Given one sentence in the source language, this input will go through all those components until the system outputs a set of pairs mapping candidate translations to probabilities. The Stat-XFER framework is shown in the figure 1 below.

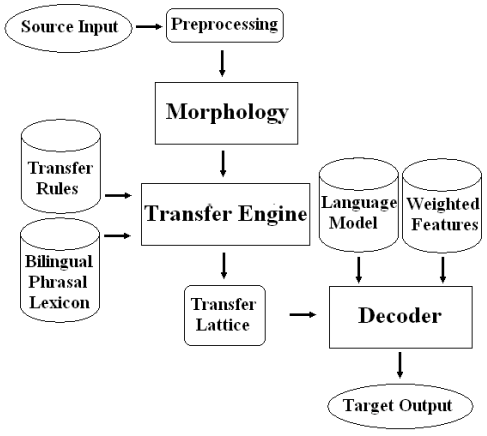

Figure 1. The Stat-XFER framework

Proceedings of the NAACL HLT 2010 Student Research Workshop, pages 58–63,
Los Angeles, California, June 2010.

Bilingual Phrasal Lexicon. The Portuguese-English lexicon is made up of a word-level lexicon. The lexicon was automatically extracted from the EUROPARL corpus[1] and in order to adjust it to this system's needs, this lexicon has been modified with new entries and deleted of the repeated entries.

Morphology. The system uses the Jspell[2] morphology analyzer, which was developed by the Minho University in Portugal. Jspell produces all possible labels for a lexical entry in the input and it provides a stem for each word with a different label. Upon each sentence in the lexicon input, the Morphology Analyzer examines each lexical entry and performs its analysis.

Transfer Engine. This component was developed by the AVENUE group from Carnegie Mellon University (Probst et al., 2002, Lavie et al., 2003, Lavie et al., 2004, Lavie 2008) and applies the lexical transfer rules specified by the Lexicon component, i.e. the lexical entries in Portuguese are substituted by their translations in English. This component also applies the transfer grammar rules from Portuguese into English producing constituents translated into the English structure. When the transfer stage has completed, we have a chart (hypergraph) in which each key contains an n-best beam for each non-terminal type of each source span. Each entry in the beam has a unification feature structure and log-linear feature scores associated with it.

Before decoding, the chart is turned into a lattice by removing hierarchical structure and respecting only the target-side ordering of the constituents produced by the grammar. The unification features and unification constraints on each rule can remove ungrammatical hypotheses from the search space.

Decoder. For each sentence, the Decoder does a monotonic left-to-right decoding on the Translation Lattice with no reordering beyond what the grammar produced. For each source span, the decoder keeps the *n* most likely hypotheses.

3 Experiment

According to Ma and McKeown (2009), in translations the main verb of a sentence is the most important element for the sentence comprehension. In Machine Translation, the goal is to produce understandable translations. Therefore, we started our experiment by adjusting the verbal tense rules between the Brazilian Portuguese and English.

Natural languages do not share the same properties concerning verbs. The pair of languages Portuguese and English for example, has different systems of modality. According to Palmer (1986), the English language has its system of modal verbs defined by *can*, *could*, *will*, *would*, *may*, *must*, *might*, *ought*, *shall*, *should*, *need* and *dare*. However, the Portuguese language has a system of mood consisting of *indicative* and *subjunctive* moods (Bechara, 2002).

Moreover, the Brazilian Portuguese and the English languages present morphological differences between verbal forms. For example, many verbs in English have the following form: Base + {-s form (3rd person singular present)/PAST/-ing/-ed} (Quirk and Greenbaum, 1973). Nonetheless, in Brazilian Portuguese the verbs present the following form in most case: Base + thematic vowel + number/person agreement + tense/mood agreement. This verbal form is present in every tense in the *indicative* and *subjunctive* moods.

In order not to lose any information concerning the distinction between verbal tenses from Brazilian Portuguese to English, we built a corpus with sentences in all verbal tenses in Portuguese (Bechara, 2002) and manually mapped them into English. Each Portuguese verbal form was mapped into English in all their conceivable translations. This corpus is a sentence-level parallel corpus with original sentences in the Portuguese language and their respective translations. The main goal of building this corpus was to verify all possible translations between Portuguese and English verbal tenses.

3.1 Methodology

Since this was a pilot experiment, we were interested in verifying if our changes improved the system. This pilot experiment consists of translating a corpus containing all three verb conjugation classes in Portuguese: 1) First conjugation class: verbs ending in –AR; 2) Second conjugation class:

[1] http://www.statmt.org/europarl/
[2] http://linguateca.di.uminho.pt/webjspell/jsolhelp.pl

verbs which end in –ER; and 3) Third Conjugation class: verbs ending in –IR, respecting each tense, mood and number in Stat-XFER system and then evaluating their results.

To construct the three corpora – one corpus for each conjugation class – we extracted the 100 most frequent Portuguese verbs appearing in Google's search engine[3] in all three conjugation classes in Portuguese. The search for the most frequent Portuguese verbs was done using a tool developed in Python and followed the following constrains: a) a result had to be an infinitive verb; b) had to be in the Portuguese language and c) had to be found in a Brazilian Web page.

From the one hundred most frequent Portuguese verbs in each conjugation class we manually built three corpora of all three verb conjugation classes in Portuguese, respecting each tense, mood, number, and person, and then a human translator mapped them into English through manual translation. Each corpus contains no more than three examples of each most frequent verb selected, resulting in 163 sentences.

Once all verbal tenses were translated into English, we applied these three corpora to the Stat-XFER system and evaluated all resulting translations using Meteor. Meteor is a metric for the evaluation of Machine Translation output. The metric is based on the harmonic mean of unigram precision and recall (Lavie and Agarwal, 2007). The Meteor scores are in a scale from 0 to 1, where 0 means the translation is the farthest from the reference translation and 1 means the translation is most similar to a human translation.

After the evaluation, we initiated the improvement of the grammar module with new syntactic transfer rules in order to deal with the problems presented by the differences between the verbal tenses. The transfer rules were manually developed and encoded how constituent structures in the source language transfer to the target language. In the beginning of our research, the system had 113 such rules in its grammar, but with some modifications the system now has 152 rules concerning the mapping from Portuguese to English. We add a rule for each tense in each conjugation class. Few rules address more than one tense. A Stat-XFER syntactic rule example is shown below.

```
1 {VP, 2}
2 ;;SL: ANDO
3 ;;TL: WALK
4 VP::VP [V] → [V]
5 (
6  (X1::Y1)
7  ((X1 tense) = c pres)
8  ((X1 mood) =c (*NOT* subj))
9  ((Y1 tense) = pres)
10 ((X0 number) = (X1 number))
11 ((X0 person) = (X1 person))
12 )
```

The first line is the name of rule, in this case *VP, 2*. The second and third lines are examples of transference between both languages, which means a source language (*SL*) will be encoded into a respective target language (*TL*). The forth line indicates that a simple verb in *SL* will be translated as a simple verb in *TL* as well.

The condition of application of that rules is between parenthesis, shown in the fifth and twelfth lines. In line six, inside these parentheses, it is indicated that the first element of the verbal phrase from source language, i.e. *X1*, will be converted as the first element of the verbal phrase in the target language, i.e. *Y1*. Line seven says that *X1* must be in the present tense and line eight says it must not be a subjunctive mood. Line nine states that *Y1* must be in the present tense. Lines ten and eleven indicate that *X1* will receive the number and person from an element in the source language, *X0* in this case.

Another significant modification concerns the Bilingual phrasal lexicon. In the beginning of our research, the system had 56.665 entries with their respective translations. However, many of these lexical items presented improper translations, repetitions and lack of correct meanings. To help improve the development of our system some modifications in this lexicon were required. This word level lexicon is in constant modification with the inclusion of missing lexicon items, cleaning of repeated items and correction of the inappropriate ones. The Stat-XFER word-level lexicon has now 57.315 entries with their respective translations.

3.2 Results

After the insertion of new syntactic rules, adjustment of the existing (old) rules and the modifica-

[3] http://www.google.com.br

tions in the lexicon, we translated the same three corpora again to evaluate the performance one more time. The comparative results of these two evaluations are shown in Table 1 and Figure 2 below.

Corpora	System in the initial state	System in the current state
1st conjugation class	0.5346	0.5184
2nd conjugation class	0.5182	0.5269
3rd conjugation class	0.5291	0.5356

Table 1. Meteor evaluation of initial and current state of the system

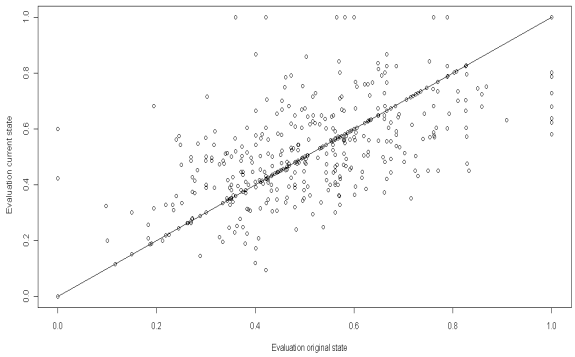

Figure 2.The graphic shows evaluation results of a sample of 489 sentences distributed across all verbal tense, mood, number and person in English during the initial stage and the current stage of the development of the Stat-XFER system. Results are in a scale of 0 to 1. 0 means the translation is the farthest from the reference translation and 1 means the translation is most correlated to a human translation. Each entry in the picture corresponds to one sentence evaluated according to the Meteor metric, distributed in the x-axis (initial evaluation) and the y-axis (current evaluation). The 45o line distinguishes those cases where performance of one evaluation was better than the other.

Interesting results were collected in this experiment. The new results show improvements in correlation with translation references on the second and third conjugations (Table 1). The system modifications we have done so far yielded some improvements. The inclusion and correction of syntactic rules according to the differences between linguistic parameters of the languages is the natural way to improve the results of Meteor evaluation in Stat-XFER system.

However, it is also noticeable that in the first conjugation we observed a decline in the performance of the system. This was one of the interesting observations we made based on this study.

To better study the effect of the inclusion of new grammar rules between verbal tenses, we perform a second evaluation of our system, now using a previously unseen corpus. This new corpus was build from FAPESP Magazine[4], which is a bilingual online publication designed for the Brazilian scientific community. We extracted 415 sentences from the Humanities section and translated them with the old and new syntactic rules to evaluate the performance of the system. The comparative results of these two evaluations are shown in Table 2 below.

FAPESP Corpus	System in the initial state	System in the current state
Humanities section	0.2884	0.5565

Table 2. Meteor evaluation of initial and actual state of the system

The two evaluations of the FAPESP Magazine corpus indicate that the insertion of new grammar rules presents significant improvement compared to the system in its initial state. It is important to note that these results validate our experiment and confirm the improvement of the system. We discuss our results in Section 4.

4 Discussion

An interesting observation is that in the corpus manually built for the experiment, the score of the first conjugation decreased (Table 1), while we expected that it should increase. We were surprised by this phenomenon and started investigating its causes. After preliminary studies, we now believe that this is due to a greater number of possible meanings that verbs from the first conjugation in Portuguese can assume, thus producing several translations in English, with smaller a correlation with respective human reference translations.

According to Williams (1962), the endings of infinitive verbs in Portuguese are derived from Classical Latin. The first conjugation in Brazilian Portuguese also contains verbs borrowed from dif-

[4] http://revistapesquisa.fapesp.br

ferent languages, e.g. the English verb *to delete* has its correspondent *deletar* in Portuguese. Moreover, the creation of new verbs in Portuguese is always included in the first conjugation, sharing a common set of characteristic of verbs ended in –AR. While the second and third conjugation classes tend to have a finite number of verbs given from Classical Latin, the first conjugation class is still increasing, unlike the second and third conjugations. These changes in verbs from the first conjugation classes may impact the development and evaluation of the system.

Note that the corpora used to evaluate our system are very small compared to corpora recommend for Machine Translation evaluations. Since this was a pilot experiment, we were only interested in verifying if the questions we wanted to ask were answerable. Further validation will be performed once more rules are added to the system and new human translations are included in the reference corpus.

Lavie et al. (2004) applied a transfer-rule learning approach to this system in order to learn automatically the transfer rules from Hebrew to English. We believe that this approach can be applied to the Portuguese-to-English system and it can improve the coverage of grammar rules.

5 Conclusion

The results of our experiment are very promising. We could observe a clear improvement in the performance of the system after syntactic rules were added to its grammar module. The new syntactic rules concerning the verbal tenses improved system performance, but also indicated that there is significant room for improvements in the Stat-XFER system. In particular, improving the transfer rules in other aspects beyond the Verbal tense is a promising area of future research.

Although this research is in a preliminary stage, we already made interesting linguistic observations about Portuguese verbs from first conjugation concerning the development of Machine Translation systems. We believe this is a general issue that should concern every designer of Machine Translation system of the Brazilian Portuguese language. We are very excited about the future stages of this study, and its potential contribution to the linguistic perspective of the field of Machine Translation.

Acknowledgments

We would like to thank Fidel Beraldi and Indaiá Bassani from University of São Paulo, for developing the Python tool, which extracts the frequency of the verbs from Google search engine; Bianca Oliveira from Primacy Translations for translating the corpus reference; Jonathan Clark from Carnegie Mellon University for helping the authors with clarifications about the system; Marcello Modesto from University of São Paulo, for advisorship and Alon Lavie from Carnegie Mellon University, for hospitality and technical clarifications. The author acknowledges that the initial Stat-XFER system was developed jointly by Modesto, Lavie and the entire AVENUE group (www.cs.cmu.edu/~avenue).

References

Chiang, D., Knight, K. and W. Wang. 2009. 11,001 New Features for Statistical Machine Translation. *Proceedings of NAACL-HLT*.

Comrie, B. 1993. Tense. Cambridge University Press.

Ma, W., McKeown, K. 2009. Where's the Verb? Correcting Machine Translation During Answering. *Proceedings of the ACL-IJCNLP*.

Bechara, E. 2002. *Moderna gramática do Português*. Lucerna.

Lavie, A., Vogel, S., Levin, L., Peterson, E., Probst, K., Font Llitjos, A., Reynolds, R., Carbonell, J., and Cohen, R. 2003. Experiments with a Hindi-to-English Transfer-based MT System under a Miserly Data Scenario. *ACM Transactions on Asian Language Information Processing* (TALIP), 2(2).

Lavie, A., Wintner, S., Eytani, Y., Peterson, E. and Probst, K. 2004. Rapid Prototyping of a Transfer-based Hebrew-to-English Machine Translation System. *Proceedings of the 10th International Conference on Theoretical Methodological Issues in Machine Translation* (TMI-2004).

Lavie, A., Agarwal, A. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. *Proceedings of the 2nd Workshop on Statistical Machine Translation at the 45th Meeting of the ACL (ACL-2007),* pp. 228—231.

Lavie, A. 2008. Stat-XFER: A General Search-based Syntax-driven Framework for Machine Translation. Invited paper in Proceedings of CICLing-2008.

Haifa, Israel, February 2008. Gelbuch (ed.), *Computational Linguistics and Intelligent Text Processing*, LNCS 4919, Springer. pp. 362-375.

Palmer, F. 1986. *Mood and Modality*. Cambridge University Press.

Probst, K., Levin, L., Peterson, E., Lavie, A., and Carbonell, J. 2002. MT for Minority Languages Using Elicitation-Based Learning of Syntactic Transfer Rules. Machine Translation, 17 (4).

Quirk, R., Greenbaum, S. 1973. *A University Grammar of English*. Longman.

Williams, E. 1962. *From Latin to Portuguese: Historical Phonology and Morphology of the Portuguese Language*. University of Pennsylvania Press.