Statistical Phrase-based Machine Translation: Experiments with Brazilian Portuguese

Wilker F. Aziz, Thiago A. S. Pardo [1], Ivandré Paraboni [2]

[1]NILC/ICMC, Universidade de São Paulo
Av. Trabalhador São-Carlense, 400, São Carlos, Brazil

[2]EACH, Universidade de São Paulo
Av. Arlindo Bettio, 1000, São Paulo, Brazil

wilker.aziz@usp.br, taspardo@icmc.usp.br, ivandre@usp.br

Abstract. *Statistical approaches have recently emerged as the main paradigm in Machine Translation (MT) research. In previous work we have shown that results of a simple statistical word-based MT system may be highly comparable to those produced by a rule-based approach for closely-related languages such as Brazilian Portuguese and European Spanish. In this work we take the discussion one step further and present evidence that a more sophisticated (namely, phrase-based) translation model may outperform rule-based translation for this language pair, and additional results of a first experiment in Portuguese/English phrase-based statistical MT.*

1. Introduction

Statistical Machine Translation (SMT) addresses the computational problem of translating sentences from one natural language (the 'source' language) into another (the 'target' language) by analysing large amounts of parallel text – i.e., aligned sentence pairs conveying a source sentence and its translation into the target language – and building statistical translation models from them. Among other benefits, (purely) statistical approaches require by definition no hand-coded linguistic knowledge, and are entirely trainable from corpora in any given source/target language pair.

Early SMT systems achieved interesting but limited results by training translation models from alignments at word level (e.g., Brown et al., 1993), which is known as *word-based* SMT. More recently, however, the field has witnessed significant advances in translation quality with the use of *phrase-based* SMT techniques, that is, models that take into account the alignment of arbitrary sequences of words, which may or may not be linguistically-motivated (Och et al., 2003).

Regarding the use of SMT techniques for the Brazilian Portuguese Language, in previous work we have proposed a simple word-based SMT system for the Brazilian Portuguese and Spanish languages. Results reported in Aziz et. al. (2008) suggest that our system, despite using a training data set deemed insufficient for any practical purposes, was only slightly inferior to a rule-based approach, the Apertium system (Corbí-Bellot et al., 2005).

These encouraging first results allowed us to predict that with a more sophisticated translation model and/or larger amounts of training data our system could

easily outperform the rule-based approach. In this paper we revisit this issue by presenting a series of new experiments in Brazilian Portuguese, European Spanish and now also American English machine translation, and using state-of-art phrase-based SMT models. The results of both word-based and phrase-based SMT are compared with those obtained from Apertium (Corbí-Bellot et.al., 2005) and Google Translate[1], providing further evidence in favour of SMT systems in general and, in particular, supporting the use of phrase-based models in SMT.

The rest of this paper is structured as follows. In the next section we introduce the SMT paradigm. Section 3 reports our experiments, whose results are discussed in Section 4. Section 5 presents our conclusions and suggests some future work.

2. Background

Statistical Machine Translation was first formalized by Brown et al. (1993) and then reformulated as an optimization problem by Och & Ney (2004). In general terms, a statistical approach to translate from, e.g., Spanish to Portuguese, involves finding the Portuguese sentence *p* that maximizes the probability of *p* given a Spanish sentence *e*, that is, the probability $P(p \mid e)$ of *p* being a translation of *e*. Using the Bayes theorem, we have that $P(p \mid e) = P(p)\, P(e \mid p) / P(e)$. Since *e* is given, $P(e)$ is constant and may be eliminated from the formula. Therefore, we look for the translation *p* that maximizes $P(e \mid p)\ P(p)$. The probability $P(p)$ (as well as $P(e)$) may be easily obtained from language models.

Given a pair <*e, p*> of sentences, *e* and *p* are said to be mutual translations if there is at least one possible alignment among them, i.e., correspondences among their words and/or phrases. Assuming the set of all possible alignments among *e* and *p* to be *a*, obtaining $P(e \mid p)$ can be seen as the maximization of the sum of individual contributions of every single alignment a_i, a process called decoding:

$$P(p \mid e) = P(p) \sum_i P(a_i, e \mid p)$$

Early SMT systems used information obtained directly from word alignments between each sentence pair <*e, p*>, a method known as *word-based* translation. Examples include the original IBM translation models that started the research in the field in the late 1980s using word alignments obtained with the Giza++ toolkit (Och & Ney, 2003) and A* greedy decoding.

More recently, however, a number of studies have obtained superior translation quality by examining phrase alignments, that is, arbitrary sequences of words that may or may not be linguistically-motivated. This method, known as *phrase-based* translation, appears now in many flavours in the SMT literature and represents the state of the art in the field. One of the best-known studies of this kind is the basic phrase-based translation model presented by Koehn et al., (2003). For a comparison between phrase-based and other recent (e.g., hierarchical) approaches to SMT, see Zollmann et al. (2008).

There are many ways in which phrase translation pairs are extracted. Perhaps surprisingly, some of the best results have been obtained with relatively simple

[1] http://translate.google.com/

extraction heuristics based on word alignment (Och et al., 2003), that is, the kind of information readily available as a by-product of word-based SMT tools such as Giza++ (Och & Ney, 2003). Additionally, it has been shown that lengthy (above 3 or 4-words long) or linguistically-motivated phrase models (e.g., those that consider phrases only in the standard linguistic sense) tend to perform poorly. In particular, imposing syntactic restrictions to avoid non-intuitive phrase definitions such as "house the" is shown to be harmful, the reason being that systems of this kind filter out a large number of otherwise useful phrases (Och et al., 2003).

Phrase-based decoding as proposed by Och et al., (2003) works as follows. A source input sentence (e.g., in Spanish) e is segmented into a sequence of I phrases e_1^I assuming a uniform probability distribution over all possible segmentations. Each source phrase e_i in e_1^I is translated into a (possibly reordered) target phrase p_i, and phrase translation is modelled as a probability distribution $\varphi(e_i \mid p_i)$. Reordering of the target output phrases is modelled by a relative distortion probability

$$d(start_i, end_{i-1}) = \alpha^{|start_i - end_{-1}| - 1}$$

with an appropriate value for the parameter α. To optimize performance, a word cost factor ω is introduced[2] for each generated target word in addition to the language model (LM). Thus, the best target output (e.g., Portuguese) sentence p_{best} given a source input (e.g., Spanish) sentence e is

$$p_{best} = \text{argmax}_p \, p(p \mid e) = \text{argmax}_p \, p(e \mid p) \, \text{LM}(p) \, \omega^{\text{length}(p)}$$

where $p(e \mid p)$ is decomposed into $p(e_1^I \mid p_1^I) = \Pi_{i=1}^I \, \phi(e_i \mid p_i) \, d(\text{start}_i, \text{end}_{i-1})$.

Many of the tools required for building phrase-based SMT models as above are available from the Moses toolkit (Koehn et al., 2007). These include the basic resources for training translation models from corpora in any language, and an efficient beam-search decoding algorithm. In addition to that, factored models allow the inclusion of linguistic knowledge (e.g., morphology, word classes, etc.) to the pre-processing or post-processing stages, and confusion network decoding facilities supporting the translation of ambiguous input (e.g., useful for the integration of speech recognition and machine translation.)

Moses refines word alignments produced by Giza++ by re-implementing a number of methods proposed by Och et al. (2003). The tool uses heuristic rules to collect all aligned phrases that are consistent with the word alignment, and estimates phrase translation probability distribution by relative frequency without smoothing[3]. Decoding is performed by beam search with recombination of translation hypotheses to reduce search space: hypotheses can be recombined if they agree in (a) the foreign words cover so far, (b) the last two target words generated, and (c) the end of the last foreign phrase covered. For details we refer to Koehn et al. (2007).

3. Experiments

In previous work we have proposed a simple SMT system that we have called *Z**. This consisted of the early IBM 4 translation model based on alignments at word level

[2] The value of ω is usually larger than 1, with a bias toward longer output cf. Koehn et al., (2003).
[3] Data sparseness is handled using the concept of *lexical weighting* discussed by Och et al. (2003).

(Brown et al., 1993) and it was built using Giza++ (Och and Ney, 2003), the decoding facilities provided by the ISI ReWrite Decoder tool (Germann et al., 2001) and language models created with the aid of the CMU-Language toolkit (Clarkson and Rosenfeld, 1997). For ease of comparison to our present work, we will rename this system as *Z*w*, in which 'w' stands for '*word-based*' statistical machine translation.

The experiments with *Z*w* reported in Aziz et. al. (2008) for Brazilian Portuguese and European Spanish languages, although using a very small training data set (about 17,000 sentence pairs), showed that our system was only slightly inferior[4] to the rule-based system Apertium (Corbí-Bellot et al., 2005). This allowed us to predict that with a more sophisticated translation model and/or larger amounts of training data[5] our system may outperform the rule-based approach.

Given the difficulties in obtaining large, reliably-aligned corpora for Brazilian Portuguese, in this work we further the issue by looking into alternative translation models. More specifically, we will consider state-of-art *phrase-based* translation models (e.g., Koehn et al., 2003), and we will apply them to develop a series of new translation experiments involving Brazilian Portuguese, European Spanish and also American English languages.

The purpose of our experiments is twofold: first, we would like to examine whether phrase-based SMT for Brazilian Portuguese is indeed superior to word-based SMT and, if so, by how much. Second, we would like to show that the slight advantage observed in rule-based translation as reported in Aziz et. al. (2008) is lost as we introduce phrase-based models, and that is the case even when using our unrealistically small training data set.

We used Moses (Koehn et al., 2007) and language models created using the SLRIM toolkit (Sotlcke, 2002) to developed two versions of a phrase-based SMT system, called $Z^*p_{n=3}$ and $Z^*p_{n=5}$ (in which 'p' stands for *phrase-based* SMT and 'n' represents the order of the statistical language model under consideration.) In both cases, we used the IBM 4 model to produce basic lexical alignments that were further refined using the heuristics described in Och et. al. (2003) for phrase extraction. The phrases obtained in this way were used in the model proposed by Koehn et. al. (2003), which takes into account translation probabilities, distortion limits, lexical weighting and sentence costs.

Both phrase-based models were configured as follows: the *alignment* parameter was set to use the alignment heuristics '*grow-diag-final-and*' and we used *lexical-weighting*; the *maximum phrase length* used was set to 7 words; the *reordering* parameter was left at its default value; and the *distortion limit* was set to 6 to allow for phrase reordering.

We devised two sets of experiments in both-ways translation: one for Brazilian Portuguese (BP) and European Spanish (ES), and another for Brazilian Portuguese (BP) and American English (AE). In the BP-ES and ES-BP experiments our three SMT systems were compared with the rule-based system Apertium (Corbí-Bellot et al.,

[4] As measured by BLEU (Papineni et al., 2002) and NIST (NIST, 2002) scores.

[5] For example, benefits of using large (of about 2 trillion tokens) language models are discussed in Brants et al., (2007).

2005), which is specialised in such 'closely-related' Romance languages. For BP-AE and AE-BP translations, given that the Apertium system is unable to deal with the English language, we have compared our system with another statistical approach, namely, the large-scale Google Translate system[1], even though the vast amounts of training data available to Google make the comparison, by definition, to our disadvantage.

All experiments made use of a Portuguese-Spanish-English parallel corpus taken from the Environment, Science, Humanities, Politics and Technology supplements of the on-line edition of the "Revista Pesquisa FAPESP"[6], a Brazilian magazine on scientific news. The training data for each translation task consisted of a set of about 17,000 sentences pairs. For testing purposes each experiment used 649 previously unseen sentence pairs.

4. Results

In all the experiments we compared different MT systems by measuring BLEU (Papineni et al., 2002) and NIST (NIST, 2002) scores. BLEU and NIST are widely-used evaluation metrics based on n-gram statistics, and are intended to compute the amount of common information shared between machine and human (reference) translations. BLEU differs from NIST mainly in the way that sparse n-grams (which are considered to be more informative in NIST) are weighted, but both scores were shown to correlate well with human judgments of translation quality. In both cases, the higher the score, the better the translation quality. BLEU scores range from 0 to 1, whereas the maximum NIST value depends on the size of the data set.

For translation between Portuguese and Spanish, we compared Apertium (Corbí-Bellot et al., 2005) to our *Z*w*, *Z*p*$_3$ and *Z*p*$_5$ systems. These results are shown in the following Tables 1 and 2.

Table 1. Brazilian Portuguese (BP) to Spanish (ES) translation

System	BLEU	NIST
Z*w	0.5083	10.3106
Apertium	0.5758	10.3212
Z*p$_5$	0.6281	10.7631
Z*p$_3$	0.6292	10.7821

Table 2. Spanish (ES) to Brazilian Portuguese (BP) translation

System	BLEU	NIST
Z*w	0.4907	9.4370
Apertium	0.5561	9.9622
Z*p$_5$	0.6091	10.5908
Z*p$_3$	0.6201	10.6500

For translation between Portuguese and English, we compared Google Translate to our *Z*w*, *Z*p*$_3$ and *Z*p*$_5$ systems, once again measuring BLEU and NIST scores. The results are shown in Tables 3 and 4 below.

[6] http://www.revistapesquisa.fapesp.br/

Table 3. Brazilian Portuguese (BP) to American English (AE) translation

System	BLEU	NIST
Z*w	0.1883	6.6137
Google	0.3033	7.7662
$Z*p_5$	0.3556	8.0882
$Z*p_3$	0.3558	8.0946

Table 4. American English (AE) to Brazilian Portuguese (BP) translation

System	BLEU	NIST
Z*w	0.1288	5.3993
$Z*p_5$	0.3067	7.3763
$Z*p_3$	0.3071	7.3701
Google	0.3540	7.8280

From the above results the following observations are due: first, the phrase-based results (as measured by both BLEU and NIST scores) are indeed superior to those obtained by simple word-based SMT. In fact, our previous *Z*w* system scored lowest in all experiments. Although we presently do not seek to validate these results statistically, the difference in BLEU/NIST scores is sufficiently large to suggest real improvement in standard MT research. For a discussion on the statistical significance of BLEU/NIST scores, see Zhang et. al. (2004).

Second, phrase-based SMT outperforms the rule-based Apertium system in both BP-ES and ES-BP translations, and for both 3-gram and 5-gram language models[7]. Moreover, this is so despite the fact that we are still using a very small amount of training data if compared to current standards in the field (e.g., Brants et. al., 2007).

Third, both of our phrase-based models outperform Google Translate in BP-AE translation, but not in the AE-BP direction. This is probably due to the fact that Google is designed for the English language in the first place, and most likely equipped with a much larger English language model than its Portuguese counterpart.

Finally, we notice that for the present data set $Z*p_3$ is superior to $Z*p_5$, although the difference is not significant in AE-BP translation. This may be explained by the small size of our training data and consequent data sparseness, which is more evident in higher-ordered models.

The comparison of our systems with Apertium (for Portuguese and Spanish) and with Google (for Portuguese and English) is illustrated below, in which we consider BLEU scores only. Some translation examples are shown in the appendix at the end of this paper.

[7] Once again using a standard MT evaluation technique based on BLEU/NIST scores. From a different point of view, however, (e.g., considering word-error rates in the pos-editing task) rule-based Apertium may still outperform our *Z*w* system as discussed in Aziz et. al. (2008), although this advantage may not necessarily hold for *Z*p3* or *Z*p5*.

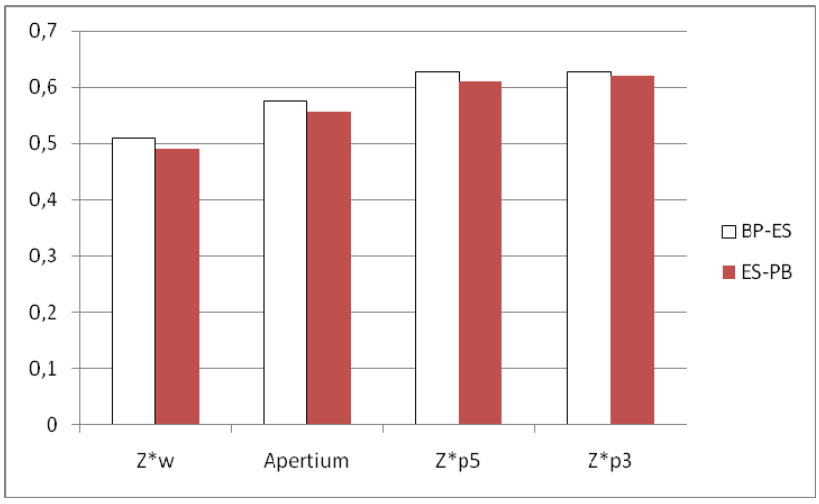

Figure 1. Portuguese←→Spanish translation using Z* and Apertium systems (BLEU scores)

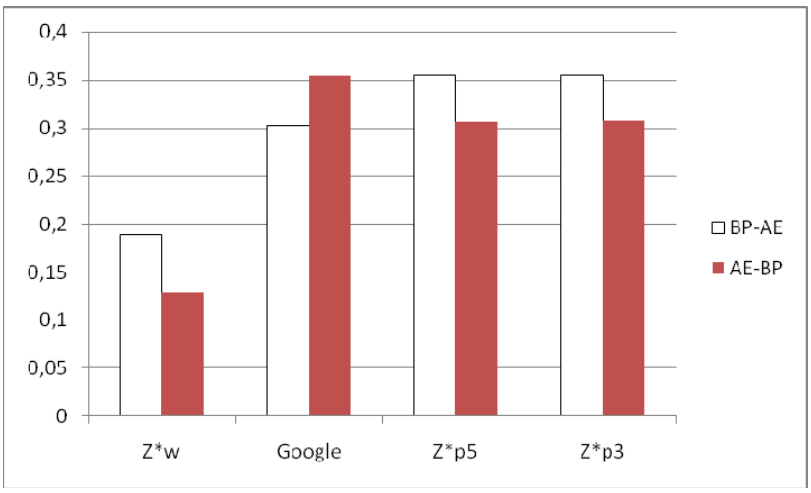

Figure 2. Portuguese ←→ English translation using Z* and Google Translate systems (BLEU scores)

5. Conclusions

In this paper we built upon our previous work and presented improved, phrase-based statistical translation models applied to Portuguese/Spanish/English machine translation. Our results show that phrase-based models may be indeed superior to our previous word-based approach. The proposed system now outperforms the rule-based Apertium System in the Portuguese/Spanish translation tasks as predicted in Aziz et. al. (2008), and preliminary results in Portuguese/English translation suggest that our system is comparable to Google Translate in the Portuguese-English direction[8] (but not so far in English-Portuguese).

[8] Although of course our system is much more specialised in the sense that they were trained on the very kind of text used as test data, whereas Google attempts to handle *any* kind of text.

As future work, we intend to improve the proposed techniques and make use of larger corpora to develop robust SMT models for Brazilian Portuguese. In addition to that, we will expand the range of language pairs under consideration by making use of the *Europarl* parallel corpus available in 11 European languages (Koehn, 2005).

Acknowledgements

The authors acknowledge support by FAPESP and CNPq.

References

Aziz, Wilker Ferreira, Thiago Alexandre Salgueiro Pardo e Ivandré Paraboni (2008) An Experiment in Portuguese-Spanish Statistical Machine Translation. 19th Brazilian Symposium on Artificial Intelligence (SBIA-2008). LNAI vol. 5249, pages 248-257. Springer-Verlag Berlin Heidelberg.

Brants, Thorsten; Ashok C. Popat; Peng Xu; Franz J. Och and Jeffrey Dean (2007) "Large language models in machine translation". The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-2007), June 28-30, Prague, pages 858-857.

Brown, P. E.; S. A. D. Pietra; V. J. D. Pietra and R. L. Mercer (1993) "The Mathematics of Statistical Machine Translation: Parameter Estimation". Computational Linguistics, Vol. 16, N. 2 pages 79-85.

Clarkson, P. R. and R. Rosenfeld (1997) "Statistical Language Modeling Using the CMU-Cambridge Toolkit". Proceedings of ESCA Eurospeech.

Corbí-Bellot, A.M.; M. L. Forcada; S. Ortiz-Rojas; J. A. Pérez-Ortiz; G. Ramírez-Sánchez; F. Sánchez-Martínez; I. Alegria; A. Mayor and K. Sarasola (2005) "An open-source shallow-transfer machine translation engine for the romance languages of Spain". 10th Annual Conference of the European Association for Machine Translation, pages 79-86.

Germann, U.; M. Jahr; Kevin Knight; Daniel Marcu and K. Yamada (2001) "Fast Decoding and Optimal Decoding for Machine Translation". Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics.

Koehn, Philipp (2005) "Europarl: A Parallel Corpus for Statistical Machine Translation". MT Summit.

Koehn, Philipp et. al. (2007) "Moses: Open Source Toolkit for Statistical Machine Translation". Annual Meeting of the Association for Computational Linguistics.

Koehn, Philipp; Franz Josef Och, and Daniel Marcu (2003) "Statistical phrase-based translation". HLT-NAACL-2003, pages 48-54.

NIST (2002) "Automatic Evaluation of Machine Translation Quality using n-gram Co-occurrence Statistics". *http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf*

Och, F.J. and H. Ney (2003) "A Systematic Comparison of Various Statistical Alignment Models". Computational Linguistics, Vol. 29, nro.1, pages 19-51.

Och, F.J. and H. Ney (2004) "The Alignment Template Approach to Statistical Machine Translation". Computational Linguistics, Vol. 30, nro.4, pages 417-449.

Papineni, K.; S. Roukos; T. Ward and W. Zhu (2002) “BLEU: a Method for Automatic Evaluation of Machine Translation”. 40th Annual Meeting of the Association for Computational Linguistics, pages 311-318.

Stolcke, A. (2002) “SRILM -- An Extensible Language Modeling Toolkit”. International Conference on Spoken Language Processing, vol. 2, Denver, pages 901-904.

Zhang Y., S. Vogel and A. Waibel (2004) “Interpreting BLEU/NIST Scores: How Much Improvement Do We Need to Have a Better System?” 4th International Conference on Language Resources and Evaluation (LREC), Lisbon, pages 2051-2054.

Zollmann, Andreas; Ashish Venugopal; Franz Och and Jay Ponte (2008) "A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT". 22nd International Conference on Computational Linguistics, pages 1145-1152.

Appendix – Sample Translations

Table 5. Portuguese source sentences (s) followed by Spanish reference (r), statistical Z^*p_3 (z) and rule-based Apertium (a) machine translations.

(s) *mas aí testaremos outro até conseguirmos o resultado desejado*
(r) si fuera así probaremos con otro hasta lograr el resultado deseado
(z) pero entonces testaremos otro hasta logremos el resultado deseado
(a) pero ahí probaremos otro hasta conseguir el resultado deseado
(s) *na região vivem cerca de 40 mil pessoas em nove favelas e três conjuntos habitacionais*
(r) en esa región viven cerca de 40 mil personas en nueve favelas y tres conjuntos habitacionales
(z) en la región viven alrededor de 40 mil personas en nueve favelas y tres conjuntos habitacionais
(a) en la región viven cerca de 40 mil personas en nueve favelas y tres conjuntos habitacionales

Table 6. Spanish source sentences (s) followed by Portuguese reference (r), statistical Z^*p_3 (z) and rule-based Apertium (a) machine translations.

(s) *si fuera así probaremos con otro hasta lograr el resultado deseado*
(r) mas aí testaremos outro até conseguirmos o resultado desejado
(z) se fora assim probaremos com outro até conseguir o resultado desejado
(a) se fosse assim provaremos com outro até conseguir o resultado desejado
(s) *en esa región viven cerca de 40 mil personas en nueve favelas y tres conjuntos habitacionales*
(r) na região vivem cerca de 40 mil pessoas em nove favelas e três conjuntos habitacionais
(z) nessa região vivem cerca de 40 mil pessoas em nove favelas e três conjuntos habitacionales
(a) nessa região vivem cerca de 40 mil pessoas em nove favelas e três conjuntos habitacionais

Table 7. English source sentences (s) followed by Portuguese reference (r), statistical Z^*p_3 (z) and Google (g) machine translations.

(s) *then we will test another until we can manage to achieve the desired result*
(r) mas aí testaremos outro até conseguirmos o resultado desejado
(z) então nós vai testar outra até podemos para conseguir surte resultado
(g) então vamos testar um outro até que possamos gerir para atingir o resultado desejado
(s) *in the region close to 40,000 people live in nine shanty towns and in three housing projects*
(r) na região vivem cerca de 40 mil pessoas em nove favelas e três conjuntos habitacionais
(z) na região cerca de 40 mil pessoas vivem em nove shanty cidades e em três habitação projetos
(g) na região perto de 40.000 pessoas vivem em favelas e nove em três projectos habitacionais

Table 8. Portuguese source sentences (s) followed by English reference, statistical Z^*p_3 (z) and Google (g) machine translations.

(s) *mas aí testaremos outro até conseguirmos o resultado desejado*
(r) then we will test another until we can manage to achieve the desired result
(z) but there testaremos another until we manage the desired result
(g) but it can test other until the desired result
(s) *na região vivem cerca de 40 mil pessoas em nove favelas e três conjuntos habitacionais*
(r) in the region close to 40,000 people live in nine shanty towns and in three housing projects
(z) in the region live some 40 1,000 persons in nine shantytowns and three sets habitacionais
(g) living in the region around 40 thousand people in nine favelas and housing three sets