Statistical Machine Translation: little changes big impacts

Helena de Medeiros Caseli[1], **Israel Aono Nunes**[1]

[1]Department of Computer Science
Federal University of São Carlos (UFSCar)
Rod. Washington Luís, Km 235, CP 676
13565-905 São Carlos-SP

helenacaseli@dc.ufscar.br, israelaono@gmail.com

Abstract. *In this paper we describe some experiments carried out to test the impact of automatic casing and punctuation changes when training and testing statistical translation models. The experiments described here concern the translation from/to English and Brazilian Portuguese texts but since the superficial changes investigated are language independent, we believe that the conclusions can be applied to many other pairs of languages. These experiments were designed aiming at setting a baseline scenario for future training and testing of more complex statistical translation models such as the factored ones. From the experiments presented here it is possible to see that case and punctuation changes have a significant impact on automatic translation results.*

1. Introduction

Machine Translation (MT) is one of the oldest and most important areas of Natural Language Processing (NLP). Since its beginnings several methods and paradigms have been proposed ranging from the basic level — in which MT is performed by just replacing words in a source language by words in a target language — to more sophisticated ones — which rely on manually created translation rules (Rule-based Machine Translation) or automatically generated statistical models (Statistical Machine Translation).

According to the automatic evaluation measures BLEU [Papineni et al. 2002] and NIST [Doddington 2002], the phrase-based statistical MT (SMT) systems such as [Koehn et al. 2003] and [Och and Ney 2004] are the state-of-the-art in MT and also the most promising paradigm for machine translation available nowadays. SMT has been widely employed since it can be applied to perhaps almost any language pair and corpus type. In fact, SMT is an inexpensive, easy and language independent way for detecting recurrent phrases that form the language and translation models.

The statistical process of translation is based on the building of two statistical models: a language model and a translation model. These models are built from a training parallel corpora (a set of source sentences and their translation into the target language) by means of IBM models [Brown et al. 1993] which calculate the probability of a given source word (or sequences of words) be translated to a target word (or sequence of words).[1] Although there is a general belief that bigger training corpus gives better translation results, it is possible to achieve satisfactory results from a training corpus of less than one million tokens in each language as shown in this paper.

[1]In SMT, the sequences of words are called phrases even though they are not really syntactic phrases.

Thus, the experiments described here were carried out to investigate some pre-processing configurations when training and testing SMT models for translating from/to English (`en`) and Brazilian Portuguese (`pt`). More specifically, these experiments aim at investigating the impact of automatic casing and punctuation changes on the statistical translation performance. The major goal is to set a baseline scenario for SMT between these languages for performing more complex statistical translation training and test in a near future. Although the experiments presented in this paper regard the SMT, it is worth mention that the relative effects of casing and punctuation changes may hold for other MT approaches such as rule-based or hybrid ones as well.

The remainder of this paper is structured as follows. Section 2 brings a briefly description of some works on SMT. Section 3 describe the experiments carried out to evaluate the impact of automatic casing and punctuation changes in the SMT performance based on the well known automatic measures BLEU and NIST. Finally, section 4 finishes this paper with some conclusions and proposals for future work.

2. Statistical Machine Translation

The SMT paradigm relies on the probabilities of source and target words (or sequences of words) to find the best translations. The background papers on this subject [Brown et al. 1993, Och and Ney 2004] describe the statistical translation process as: given a source string s, the SMT process tries to find the target translation t that maximizes the probability $p(s|t)$. This probability $p(s|t)$ is calculated as

$$p(s|t) \propto p(s)p(t|s)$$

using the Bayes Theorem, where the *translation model* $p(t|s)$ is the probability that the target string t is the translation of the source string s, and the *language model* $p(s)$ is the probability of seeing that source string. These probabilities are calculated based on the lexical alignment a_i, usually a combination between the alignment in both directions: source–target and target–word [Och et al. 2003]:

$$p(s|t) = p(s) \sum_i p(a_i, t|s)$$

The lexical alignments are generated based on IBM models [Brown et al. 1993] and HMM [Vogel et al. 1996] implemented, for example, in the automatic statistical aligner GIZA++[2] [Och and Ney 2000]. GIZA++ is already included into the open-source toolkit for SMT Moses[3] [Koehn et al. 2007] used in the experiments described here. Moses performs a phrase-based statistical translation based on the noisy channel model and the Bayes theorem rewrited as:

$$argmax_s p(s|t) = argmax_s p(t|s)p(s)$$

The target input sentence t is segmented into a sequence of n phrases with a uniform probability distribution over all possible segmentations. Each target phrase t_i is translated into a source phrase s_i which may be reordered. As mentioned before, the phrase translation is modeled by a probability $p(t_i|s_i)$, that is, due to the Bayes theorem the translation direction is inverted.

Several parameters can be used when training the statistical models using Moses

[2] `http://code.google.com/p/giza-pp/`

[3] `http://www.statmt.org/moses`

such as the GIZA++ configuration (for example which IBM models are used and how many iterations of each one has to be executed) and if the reordering has to be performed.

Although it is considered the state-of-the-art MT paradigm, the statistical models learned by means of SMT approach have the well known problem of not being able to deal with hierarchical and syntactic aspects of languages [Kitamura 2004]. Aiming at overcoming this bottleneck, some recent works are concerned with enriching the pure statistical models with syntactic, morphology and other kinds of information [Lee 2004, Sadat and Habash 2006, Och et al. 2004, Goldwater and McClosky 2005, Yamada and Knight 2001, Koehn and Hoang 2007].

From these initiatives we are particularly interested in the factored translation, which can be performed by means of Moses. Following this approach, the statistical models are learned for each level of information (lemmas, part-of-speech, morphology, syntax, etc.) together with the surface forms. According to [Koehn and Hoang 2007], the main difference between phrase-based SMT and the factored one is in the training data building and in the kind of statistical models learned from training data. In fact, phrase-based models are a special case of factored translation.

In this paper we use Moses to investigate the impact of automatic casing and punctuation changes in the performance of the SMT according to the well known automatic measures BLEU and NIST. Our purpose here is to define a baseline configuration for preprocessing training and test corpora to be able to try more complex SMT models such as the factored translation.

3. Experiments and Results

The experiments described in this paper were carried out using a corpus of 17,397 pairs of `pt-en` parallel sentences with 1,026,512 tokens (494,391 in `pt` and 532,121 in `en`). This corpus contains articles from the on line version of the Brazilian scientific magazine Pesquisa FAPESP[4] written in Brazilian Portuguese (original) and English (version).

Aiming at defining a baseline scenario for SMT `en-pt-en`, several statistical models were trained using the same Moses configuration but automatically changing casing and punctuation in training and test corpora. The main parameters of the Moses configuration were: 5 iterations of IBM-1 and HMM and 3 iterations of IBM-3 and IBM-4 for GIZA++, the maximum phrase length set to 7 and the option of reordering set as true.

Four experiments were designed regarding automatic casing and punctuation changes performed in the parallel sentences used for training and testing the translation models. Case and punctuation were selected to be investigated here since, in previous experiments, we have noticed that changing these factors when training and testing SMT models led to different results. Although simple, these superficial changes prove to significantly impact MT results as shown by the experiments presented in this paper.

Furthermore, the rationale behind the investigation of these factors relies on the learning nature of SMT methods which states that translation is learned based on recurrent patterns. The assumption behind the experiments carried out in this paper is that changing all words to lower case can improve their occurrence frequency and, hence, the

[4]Pesquisa FAPESP is available at `http://revistapesquisa.fapesp.br`.

Table 1. Experiments description regarding what is in the parallel sentences

	punctuation marks	only lower case words
E1	NO	YES
E2	NO	NO
E3	YES	YES
E4	YES	NO

Table 2. BLEU and NIST values for Portuguese (`pt`) and English (`en`) translation according to casing and punctuation changes applied in both training and test corpora

	pt–en		en–pt	
	BLEU	NIST	BLEU	NIST
E1	0.3624	8.2096	0.3247	7.6220
E2	0.3523	8.0838	0.3095	7.4354
E3	0.3903	8.3008	0.3589	7.8312
E4	0.3826	8.1971	0.3485	7.6656

MT performance. The removal of punctuation marks from training and test corpora, on the other hand, was performed to test if they have any impact in MT performance.

The features of each experiment are summarized in table 1. In the first experiment (E1), all punctuation marks were removed and also all the words were converted to lower case. In the second one (E2), only the punctuation marks were removed while in the third (E3), just the lower casing process was carried out with all the words. Finally, we also tested the models trained without applying any change in the parallel sentences of training and test corpora (E4).

Table 2 shows the values of BLEU [Papineni et al. 2002] and NIST [Doddington 2002] for the translation performed according to the automatic changes described in table 1 performed on both training and test corpora. From the BLEU and NIST values on table 2 it is possible to notice that the lower casing process and the presence of punctuation marks can improve BLEU and NIST values. For example, when only lower case words are considered in training and test corpora (E1 and E3) BLEU and NIST values are bigger than when no casing change is performed (respectively, E2 and E4). We also obtained better BLEU and NIST values when the SMT models were trained and tested with corpora containing punctuation marks (E3 and E4) than without them (respectively, E1 and E2). Summarizing, the best BLEU and NIST values were achieved when training and testing were performed considering the presence of punctuation marks and only lower case words (E3).

We also tested the impact of changing only the training corpus, that is, the changes specified in E1, E2 and E3 were applied just in the training corpus while the test one remained the same without any changes (with the punctuation marks and all words with their original upper case letters). These new versions were called, respectively E1', E2' and E3' and the values of BLEU and NIST for them are shown in table 3.[5]

[5]Since E4 was designed to do not apply any change in the corpus, the results from E4 and E4' are the

Table 3. BLEU and NIST values for Portuguese (pt) and English (en) translation according to casing and punctuation changes applied in just the training corpus

	pt–en		en–pt	
	BLEU	NIST	BLEU	NIST
E1'	0.3029	7.0192	0.2903	6.8974
E2'	0.2326	6.5197	0.3301	7.5923
E3'	0.3380	7.4885	0.2957	6.9309

Table 4. BLEU and NIST values for Portuguese (pt) and English (en) translation considering the variations of E3

	pt–en		en–pt	
	BLEU	NIST	BLEU	NIST
E3	0.3903	8.3008	0.3589	7.8312
E3'	0.3380	7.4885	0.2957	6.9309
recase	0.3053	7.0531	0.2689	6.6323

From values in table 3 it is possible to conclude that if the training corpus is changed, the test one has also to be changed. For example, for en–pt, while E3 presented a BLEU value of 0.3589, E3' gave a much worse BLEU value of 0.2957. The only case where applying the changes in the training corpus but not in the test one produced a better result than applying the changes in both was in en–pt E2'. These results show that, as expected, the SMT approach is not able to deal with case changes without being previously trained (E1 $>$ E1' and E3 $>$ E3'). Furthermore, the absence of punctuation marks in training is not a big problem since it can translate punctuation marks from source text to target text by means of a direct copy-and-paste process — sometimes it can even damage MT performance as in E2 en–pt.

The final configuration applies the Moses *recase* process in the translated sentences, that is, the same case found in the source sentence is automatically applied to the automatic translated sentence. To be able to perform this *recase* process, Moses learns (by means of a *recase* model) how to *recase* based on the original case in the source sentences. This feature, although very useful, is error-prone as can be noticed by the values of BLEU and NIST shown in the last row of table 4. This last experiment was carried out following the guidelines of Moses[6]: changing the case of all words in training and test corpora to lower case and, after the translation, applying the *recase* script using the *recase* model previously trained with the original source corpus.

An example of how translation was performed following the configuration with the best values of BLEU and NIST (E3) is shown in table 5 for pt–en and in table 6 for en–pt. In these tables we show pairs of source and translated (output) sentences together with the reference used for calculating BLEU and NIST. It is also shown the source, SMT output and reference sentences for E3' and *recase*.

same and so they are not presented in table 3.

[6]The step-by-step guidelines to train, use and evaluate Moses can be found at: http://www.statmt.org/moses_steps.html.

Table 5. Example of a Brazilian Portuguese sentence translated to English

E3	source	o centro de pesquisa de vacinas do instituto nacional de saúde dos estados unidos conseguiu a primeira vitória contra o vírus ebola .
	output	the research center of vaccines of the national institutes of health of the united states , has managed to the first victory against the ebola virus .
	reference	the research center into vaccines of the national institute of health of the united states managed the first victory against the ebola virus .
E3'	source	O Centro de Pesquisa de Vacinas do Instituto Nacional de Saúde dos Estados Unidos conseguiu a primeira vitória contra o vírus Ebola .
	output	O Centro of Pesquisa of Vacinas of the Instituto Nacional of Saúde of the Estados Unidos managed to the first victory against the virus Ebola .
	reference	The Research Center into Vaccines of the National Institute of Health of the United States managed the first victory against the Ebola virus .
recase	source	O Centro de Pesquisa de Vacinas do Instituto Nacional de Saúde dos Estados Unidos conseguiu a primeira vitória contra o vírus Ebola .
	output	The **research center** of **vaccines** of the **national institutes** of **health** , of the **united states** , has managed to the first victory against the **ebola** virus .
	reference	The Research Center into Vaccines of the National Institute of Health of the United States managed the first victory against the Ebola virus .

As can be noticed from table 5, in `pt-en` SMT, the output obtained according to E3 is just a little different from the reference sentence (just the five underlined tokens). On the other hand, the translation output by the E3' version, that is, the training with lower case but test with the original cases, led to fourteen different tokens. As expected, the E3' model was not able to translate the upper case words since the training was performed only with the lower case version of them. Finally, it is worth noticing that although the *recase* configuration generated the same mistakes as the E3 (see the underlined tokens) its values of BLEU and NIST were lower since they are calculated based on the surface forms. Thus, for example, the words *center* and *Center* in the *recase* output and the reference, respectively, are not taken as the same and neither all the other bold words. The bad behavior of *recase* process decreased the number of matches between its output and the reference causing a negative impact in BLEU and NIST values, as show the lowest values in table 4. Similar errors were found when translation the `en` sentence in table 6.

Finally, it is worth mention that the statistical models trained with a small corpus of less than 1 million tokens in each language led to satisfactory results as can be noticed from the translated sentences on the first rows of tables 5 and 6. The little errors found in these examples are of agreement or prepositions which do not prevent the understanding of the translated sentences by a human.

4. Conclusions and Future Work

In this paper we presented some experiments on analyzing the impact of automatic casing and punctuation changes in the values of BLEU and NIST using the SMT toolkit Moses. As expected, changing all words to lower case for training and testing SMT models indeed improve MT performance. The punctuation marks also prove to be very important when

Table 6. Example of an English sentence translated to Brazilian Portuguese

E3	source	the research center into vaccines of the national institute of health of the united states managed the first victory against the ebola virus .
	output	o centro de pesquisa em vacinas do instituto nacional de saúde dos estados unidos , conseguiu o primeiro vitória contra o vírus ebola .
	reference	o centro de pesquisa de vacinas do instituto nacional de saúde dos estados unidos conseguiu a primeira vitória contra o vírus ebola .
E3'	source	The Research Center into Vaccines of the National Institute of Health of the United States managed the first victory against the Ebola virus .
	output	The Research Center em Vaccines do National Institute de Health do United States conseguiu o primeiro vitória contra o Ebola vírus .
	reference	O Centro de Pesquisa de Vacinas do Instituto Nacional de Saúde dos Estados Unidos conseguiu a primeira vitória contra o vírus Ebola .
recase	source	The Research Center into Vaccines of the National Institute of Health of the United States managed the first victory against the Ebola virus .
	output	O **centro** de **pesquisa** em **vacinas** do **instituto nacional** de **saúde** dos **estados unidos** , conseguiu o primeiro vitória contra o vírus **ebola** .
	reference	O Centro de Pesquisa de Vacinas do Instituto Nacional de Saúde dos Estados Unidos conseguiu a primeira vitória contra o vírus Ebola .

learning SMT models, since they are part of the statistical phrases, their presence/absence prove to impact the MT performance. Moses also provides scripts to retrieve/apply the original source case to the translated version after automatic translation is performed. However, this *recase* process prove to be error-prone as could be noticed by BLEU and NIST values obtained following this approach.

From the experiments presented in this paper it is possible to state a baseline for SMT — training and testing using lower case version of all words and preserving the punctuation marks — which will be used in future experiments with more complex SMT models, for example, the factored ones.

Future work also include investigating post-editing rules to fix some of the grammatical errors output by the automatic SMT such as the agreement ones. It would also be investigated the improvement of Moses *recase* process.

Acknowledgments

We would like to thank the financial support of the PIADRD (Programa Integrado de Apoio ao Docente Recém Doutor) PUIC/UFSCar and FAPESP for its support in building the parallel corpus.

References

Brown, P. F., Pietra, V. J., Pietra, S. A. D., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19:263–311.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of the Human Language Technology Conference (HLT 2002)*, pages 128–132.

Goldwater, S. and McClosky, D. (2005). Improving statistical MT through morphological analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 676–683, Morristown, NJ, USA. Association for Computational Linguistics.

Kitamura, M. (2004). *Translation knowledge acquisition for pattern-based machine translation*. PhD thesis, Departament of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology.

Koehn, P. and Hoang, H. (2007). Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876, Prague.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Human Language Technology (HLT/NAACL 2003)*, pages 127–133.

Lee, Y. S. (2004). Morphological analysis for statistical machine translation. In *Proceedings of the Human Language Technology (HLT/NAACL 2004)*.

Och, F., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., and Radev, D. (2004). A smorgasbord of features for statistical machine translation. In *Proceedings of the 2004 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL 2004)*.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL (ACL 2000)*, pages 440–447, Hong Kong, China.

Och, F. J. and Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449.

Och, F. J., Ney, H., Josef, F., and Ney, O. H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318.

Sadat, F. and Habash, N. (2006). Combination of Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, pages 836–841, Copenhagen.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics (ACL 2001)*, pages 1–8, Toulouse, France.