Statistical Machine Translation of Croatian Weather Forecast: How Much Data Do We Need?

Nikola Ljubešić, Petra Bago, Damir Boras
Department of Information Sciences
Faculty of Humanities and Social Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
E-mail: {nljubesi,pbago,dboras}@ffzg.hr

Abstract *This research is a first step towards a system for translating Croatian weather forecast into multiple languages. This steps deals with the Croatian-English language pair. The parallel corpus consists of a one-year sample of the weather forecasts for the Adriatic consisting of 7,893 sentence pairs. Evaluation is performed by best known automatic evaluation measures BLUE, NIST and METEOR, as well as by evaluating manually a sample of 200 translations. In this research we have shown that with a small-sized training set and the state-of-the art Moses system, decoding can be done with 96% accuracy concerning adequacy and fluency. Additional improvement is to be expected by increasing the training set size.*

Keywords. statistical machine translation, Croatian language, English language, automatic evaluation, manual evaluation.

1 Introduction

Machine translation (MT) has a long history dating back to the 1940s. Its progress has been motivated by advances in computer science and artificial intelligence. Traditional MT was rule-based and has relied on various levels of linguistic analysis on the source side and language generation on the target side.

In the late 1980s the first statistical approach to machine translation (SMT) was pioneered by a group of researchers from IBM [2]. Since then, SMT has advanced from word-based to phrase-based models. At the beginning SMT relied on the source-channel model consisting of a translation and a language model. The translation model ensures that the system produces target hypotheses that correspond to the source sentence, while the language model ensures that the output is as grammatical and fluent as possible. Although early SMT models essentially ignored linguistic aspects, nowadays efforts are made to reintroduce linguistic information in both the translation and the language models [7].

Evaluating the output of a MT system is certainly not a simple task. Methods are usually divided into automatic, human and hybrid.

Automatic measures rely on reference translations of source sentences and calculate the likeness of the system output and the reference translations. Best known automatic measures are BLEU, NIST and METEOR.

BLEU [14] is the geometric mean of modified n-gram precisions for different n-gram lengths (usually from one to four), multiplied by a factor (brevity penalty) that penalizes producing short sentences containing only highly reliable portions of the translation.

NIST [5] is the arithmetic mean of clipped n-gram precisions for different n-gram lengths multiplied by a brevity penalty. Also, when computing the NIST score, n-grams are weighted according to their frequency, so that less frequent (and thus more informative) n-grams are given more weight.

While BLEU and NIST are based on precision, METEOR [1] calculates both recall and precision on the unigram level assigning in the harmonic mean more weight to recall than precision. Additionally, METEOR enables matching on the stem and synonym level. All three measures correlate highly (around 0.9) with human judgments at the corpus level. METEOR is reported to correlate higher than BLEU and NIST[1]. Another advantage of

METEOR is that it produces also scores on sentence level. However, correlation with human judgment on sentence level is much lower than on corpus level (0.403 in [1]).

Human evaluation mostly consists of scoring every translation by adequacy and fluency on a scale from 0 to 5. Adequacy indicates the extent to which the information contained in the source is included in the translation, whereas fluency measures how grammatical and natural the translation sounds.

2 Experimental design

The corpus used in this research is a one-year sample of the weather forecasts for the Adriatic published by the Croatian Meteorological and Hydrological Service [4]. The forecasts are published twice a day in four languages: Croatian, English, German and Italian.

This research deals only with the Croatian-English language pair. The pair consists of 720 documents and 2800 paragraphs (4 paragraphs/sections per document). Building the translation model and the decoding is performed with the Moses system [12]. The input for training the translation model with Moses is a sentence-aligned corpus. For this reason, the corpus is sentence split and tokenized. The Croatian part consists of 8,409 sentences and the English part consists of 8,368 sentences. Furthermore, the sentences are aligned by the Gale & Church sentence alignment algorithm [6]. For the sake of simplicity, those sentences that translate into one sentence only are chosen. The resulting sentence-aligned corpus consists of 7,893 sentence pairs. Thereby, some 6% of data is lost.

The basic questions this research deals with are:

1. the amount of data necessary for training a translation model for the level of text complexity of weather forecasts and
2. level of quality of the final translation.

The first question will be answered through an experiment where the translation and language models are trained on ten different corpus sizes. In these ten steps the corpus size, on which the models are trained and tested, grows from 789 to 7,890 sentence pairs.

To ensure a good estimate of the calculated measures, every step is repeated ten times. In each iteration, the sample is built from scratch by selecting sentence pairs randomly from the whole corpus. After building a sample, it is split into a 9:1 ratio - a training and a test set, respectively.

The training procedure consists of training the language model with the SRILM tool [15] on the English training instance using Kneser-Ney smoothing [9] which is often proven to achieve best results [3]. The translation model is trained in Moses [8] with default settings. These are defined in the script `train-factored-phrase-model.perl` [10].

After training the translation model, Croatian sentences are decoded using the default Moses settings. The output of the decoding step is evaluated by three previously described methods - BLEU, NIST (implementations mteval-v13a [13]) and METEOR (implementation meteor-1.0 [11]).

Unknown words are also monitored and recorded through the unknown word rate (UWR) which is the percentage of words in the source that have no translation in the translation model. It is important to note that the UWR measure is not an evaluation measure, since it is constant for a specific training and test set and does not depend on the machine translation method.

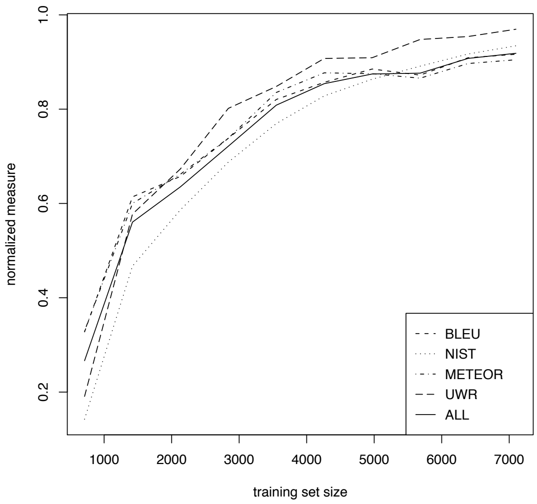
The recorded measures are used for two tasks: calculating the progress rate as the corpus size increases, and observing the correlation between the four measures.

At the end of this research, human evaluation is performed on a sample of 200 sentences. The primary goal of this evaluation is to achieve a clear insight in the quality of the translations. Error types are also recorded giving additional information regarding the causes of the observed translation errors. Results of human evaluation are compared to METEOR and UWR, since only these measures are capable of calculating agreement on the sentence level.

Table 1: Corpus statistics (wps - words per sentence, cps - characters per sentence)

language	Croatian	English
mean(wps)	10.589	13.652
mean(cps)	69.136	72.237
count(token)	87681	111944
count(type)	802	592
type-token ratio	0.00915	0.00529

Figure 1: Normalized recorded measures as the training set size increases



3 Results

The aligned corpus used in the research is described through some basic statistics in table 1. The statistics show that, as expected, Croatian has a lower token count, but a higher type count due to its rich inflectional morphology. Furthermore, English sentences tend to consist of more characters and more words. In general, the type-token ratio emphasizes the overall simplicity of the text with only 802 (Croatian), ie. 592 (English) types on approximately 100,000 tokens.

As described in the experimental design, three automatic evaluation metrics and the unknown word rate (UWR) are recorded as the corpus size increases in ten steps. On every step ten iterations are undertaken. The results are normalized to the [0, 1] scale for easy visual and numerical comparison.

Figure 2: Scatter plot of the four recorded measures

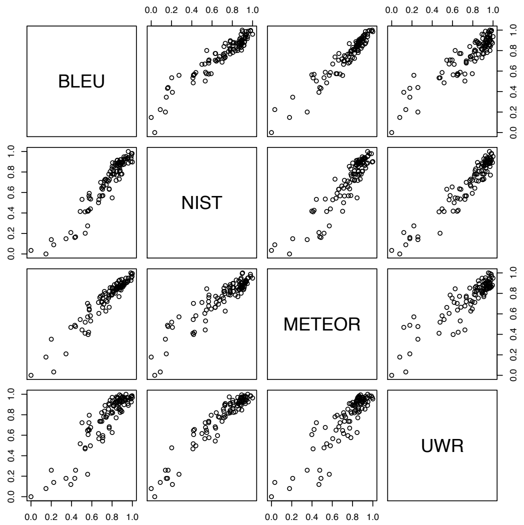


Table 2: Pearson correlation coefficient between the four measures

	NIST	METEOR	UWR
BLEU	0.95179	0.95590	0.91322
NIST		0.91960	0.95803
METEOR			0.89478

Thereby, UWR changes its sign since it grows as the translation quality decreases. Figure 1 shows the mean of the ten iterations on the four measures as corpus size increases in ten steps. The mean of all four measures is shown with a full line (ALL).

The results show an obvious consistency between different measures. Additional improvement is to be expected as the corpus size increases. Interestingly, the NIST measure has the smoothest curve showing the least sensitivity to different data.

In the next step, the relationship between the four measures is shown in a scatter plot in figure 2.

The correlation coefficients of the four variables are shown in table 2. The data show that BLEU and METEOR have the most consistent results. BLEU and NIST, in theory the most similar measures, are second most consistent, while METEOR and UWR have the lowest correlation coefficient.

Table 3: Standard deviation of results of the four measures in the final step

measure	standard deviation
BLEU	0.053
NIST	0.037
METEOR	0.045
UWR	0.021

Since NIST has the smoothest curve as corpus size increases, its data is used to calculate the progress on specific steps. In the last three steps, the progress is 3.2%, 2.9% and 1.9%, respectively. These numbers show, as does figure 1, the possibility of further improvement of results by increasing the corpus size.

Regarding the unknown word rate, it is on average 5.52% on the smallest corpus, whereas on the biggest corpus it drops to 0.86%. Percentages of decrease, as the corpus size increases, correspond to the previous numbers given for the NIST metric.

Additionally, standard deviation of a specific metric is calculated to examine the consistency of the results. Standard deviation is calculated on normalized results for the tenth and final step. The results for the specific measures are shown in table 3. Unknown word rate deviates least from its central tendency. NIST is the metric with the lowest standard deviation among the evaluation metrics. This corresponds to the smoothness of its curve in figure 1. BLEU is the metric with the highest deviation.

At the end of this research, human evaluation of 200 translations is undertaken to get a clear picture of the quality of the automated translation. Out of ten experiments with the largest corpus size, sampling is done from the results that contain most of the medians of the four recorded measures. Out of these 7890 translations, 200 random translations are chosen and given to the human evaluator for manual evaluation.

The human evaluator is first given the target translation to evaluate its fluency and later the source to evaluate adequacy. Both adequacy and fluency are graded on a scale

Table 4: Frequency of error types

error type	absolute	relative
type 1	23	0.397
type 2	20	0.345
type 3	10	0.172
type 4	5	0.086

Table 5: Pearson's correlation coefficient between human evaluation (HE) and the UWR and METEOR measures

	UWR	METEOR
HE	0.52325	0.22604
UWR		0.24777

from 0 to 5.

In addition, if the grade is less than 5 on any of the criteria, the error type is also recorded. There are four error types:

1. all lexical items correct, but meaning changed by word order or punctuation
2. lexical item translated incorrectly
3. unknown word in the source
4. typing error in the source

Frequency of these error types is given in table 4.

From the grades given by the human evaluator, accuracy is calculated as the percentage of the assigned grades regarding the maximum grade. The accuracy given by the human evaluation is 96.15% on the sentence level. If the length of the sentence is taken into account, accuracy drops down to 93.631%.

Since only a part of a sample was evaluated by humans, it is impossible to compare the result of human evaluation and automated evaluation on the corpus level. Out of four recorded measures, METEOR and UWR can also be calculated on the sentence level. The Pearson correlation coefficient between human evaluation (HE) and these two measures is given in table 5. Correlation between human evaluation and unknown word rate is over 0.5, while METEOR and human evaluation correlate with only 0.22.

4 Conclusion

In this research we have shown that with a small-sized sentence-aligned parallel corpus, and the state-of-the-art Moses system, decoding can be done with 96% accuracy concerning adequacy and fluency. It is important to note that this domain-specific text is very simple, having only 600 types on almost 100,000 tokens.

As corpus size increases, automatic evaluation measures behave in a typical logarithmic way. With around 7,000 sentence pairs of training data, improvement falls down to 2%. Additional training data could further improve the results.

The relationship between the automatic evaluation measures BLEU, NIST and METEOR is also explored. All these metrics correlate very highly with each other as well as with the negated UWR.

Exploring the correlation of METEOR and UWR with the human evaluation on sentence level shows a good correlation with UWR, but, as expected, a low correlation with METEOR.

The behavior of these automatic evaluation measures is still rather unknown, and we believe that this research has shed some light on it.

References

[1] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

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

[3] S. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the ACL*, pages 310–318, 1996.

[4] Prognoza vremena za Jadran, http://prognoza.hr/prognoze.php?id=jadran_h, 2010.

[5] G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[6] W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19(1):75–102, 1993.

[7] C. Goutte, N. Cancedda, M. Dymetman, and G. Foster. *Learning Machine Translation*. The MIT Press, 2009.

[8] H. Hoang, A. Birch, C. Callison-burch, R. Zens, R. Aachen, A. Constantin, M. Federico, N. Bertoldi, C. Dyer, B. Cowan, W. Shen, C. Moran, and O. Bojar. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 177–180, 2007.

[9] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 181—184, 1995.

[10] P. Koehn. *Moses - Statistical Machine Translation System*. University of Edinburgh, 2009.

[11] The meteor automatic machine translation evaluation system, http://www.cs.cmu.edu/ alavie/meteor/.

[12] Moses - statistical machine translation system, http://www.statmt.org/moses/, 2010.

[13] Information technology laboratory tools, http://www.itl.nist.gov/iad/mig/tools/.

[14] K. Papineni, S. Roukos, T. Ward, and W. jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, 2002.

[15] A. Stolcke. Srilm – an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, 2002.