Improving the Minimum Description Length Inference of Phrase-Based Translation Models

Jesús González-Rubio[1] and Francisco Casacuberta[1]

Pattern Recognition and Human Language Technology Center,
Universitat Politècnica de València, Camino de Vera s/n, 46021 Valencia (Spain)
{jegonzalez, fcn}@prhlt.upv.es

Abstract. We study the application of minimum description length (MDL) inference to estimate pattern recognition models for machine translation. MDL is a theoretically-sound approach whose empirical results are however below those of the state-of-the-art pipeline of training heuristics. We identify potential limitations of current MDL procedures and provide a practical approach to overcome them. Empirical results support the soundness of the proposed approach.

1 Introduction

Since their introduction, phrase-based (PB) models [1] have become the state-of-the-art pattern recognition approach to machine translation. However, despite their empirical success, their inference procedures still rely on a long decoupled pipeline of heuristics. Stages in the pipeline cannot recover errors made in earlier stages which forces each individual step to massively over-generate hypotheses. As a result, inferred phrasal lexicons [2] suffer of a huge degree of redundancy that penalizes the efficiency of PB systems. A clear indicator of these deficiencies [3] is, for example, the fact that PB models usually cannot generate the sentence pairs in which they have been trained in.

The *minimum description length* (MDL) principle [4] is a theoretically-sound alternative for PB inference. MDL, formally described in Section 3, is a general inference procedure that "learns" by "finding regularities" in the data. It embodies a form of Occam's Razor in which the best model for a given data is the one that provides a better trade-off between goodness-of-fit on training data and "expressiveness" of the model.

We build on the work by González-Rubio et al., [5] who described a practical MDL inference approach for PB models. Such approach, described in Section 4, is based on a greedy iterative procedure that generalizes an initial model that perfectly describes training data. The generalization procedure reduces the complexity of the initial model by segmenting sentences and merging common phrase pairs according to an MDL objective. Despite being theoretically-sound, these MDL PB models provide poor translation quality in comparison to state-of-the-art PB models. In Section 5, we study the potential reasons behind such poor empirical performance and propose an extended segmentation process to address this practical challenge.

Finally, the experiments in Section 6 compare the proposed approach against the MDL PB estimation proposed in [5], and a state-of-the-art PB estimation [2]. Results show that the proposed approach was able to boost the performance MDL PB models while inferring significantly smaller phrasal lexicons than conventional PB models.

2 Related Work

Different authors have proposed formal approaches to infer PB models, e.g. [6, 7]. In contrast to these approaches, MDL inference is automatically protected against overfitting and, despite being closely related to Bayesian inference, it does not suffer from its interpretation difficulties. In fact, MDL has a clear interpretation independently of whether or not there exists some underlying "true" distribution.

In [8], an MDL objective is used to prune out a phrasal lexicon previously estimated. In contrast, we use the MDL principle to directly estimate a PB model from a parallel corpora. Regarding [5], we expand their ideas by proposing an extended segmentation procedure that boosts the translation quality of MDL PB models.

Finally, our iterative segmentation process is similar to the recursive alignment model (MAR) [9]. Our method, however, learns a complete phrasal lexicon, whereas MAR only learns a word alignments as part of a larger pipeline.

3 The Minimum Description Length Principle

Given a set of data $\mathcal{D}$, the objective of statistical inference is to obtain the most probable model $\mathbf{\Phi}$ given the data. Such posterior probability can be decomposed as follows:

$$\Pr(\mathbf{\Phi} \mid \mathcal{D}) = \frac{\Pr(\mathbf{\Phi}) \cdot \Pr(\mathcal{D} \mid \mathbf{\Phi})}{\Pr(\mathcal{D})} \propto \Pr(\mathbf{\Phi}) \cdot \Pr(\mathcal{D} \mid \mathbf{\Phi}) \tag{1}$$

According to information theory [10], the negative logarithm of the probability of an event measures its description length. Therefore, searching for a MDL is equivalent to searching for a good probability distribution; which allows us to re-write Equation (1):

$$\mathrm{DL}(\mathbf{\Phi} \mid \mathcal{D}) = -\log \Pr(\mathbf{\Phi} \mid \mathcal{D}) \propto \mathrm{DL}(\mathbf{\Phi}) + \mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi}) \tag{2}$$

where function $\mathrm{DL}(\mathbf{\Phi}) = -\log \Pr(\mathbf{\Phi})$ measures the description length of model $\mathbf{\Phi}$, and $\mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi}) = -\log \Pr(\mathcal{D} \mid \mathbf{\Phi})$ denotes the description length of the data given the model.

Given these considerations, the goal of MDL inference [4] is to obtain the model $\widehat{\mathbf{\Phi}}$ with a shorter description length for a given data set $\mathcal{D}$:

$$\widehat{\mathbf{\Phi}} = \operatorname*{argmin}_{\mathbf{\Phi}} \mathrm{DL}(\mathbf{\Phi} \mid \mathcal{D}) = \operatorname*{argmin}_{\mathbf{\Phi}} \mathrm{DL}(\mathbf{\Phi}) + \mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi}) \tag{3}$$

According to the MDL principle, the best model $\widehat{\mathbf{\Phi}}$ is thus the one that reaches an optimal trade-off between model complexity and accuracy in the description of the data. A more detailed description of the MDL principles and methods can be found in [11].

4 MDL Phrase-Based Models

PB models translate following a generative process with three steps [1]: 1) the source sentence is divided into segments known as phrases, 2) each source phrase is translated into a target phrase, and 3) target phrases are reordered to conform the final translation. The main probability distribution is thus a phrase lexicon that describes the translation probability between source and target language phrases. Next sections describe how to estimate such phrase lexicons following an MDL objective [5].

Algorithm 1: Iterative inference procedure for MDL PB model estimation.

input : $\mathbf{\Phi}$ (initial PB model)
output : $\hat{\mathbf{\Phi}}$ (generalized PB model)
auxiliary: $\mathrm{collect}(\mathbf{\Phi})$ (Returns the set of possible segmentations of model $\mathbf{\Phi}$)
$\Delta\mathrm{DL}(\mathrm{s}, \mathbf{\Phi})$ (Returns variation in DL when segmenting $\mathbf{\Phi}$ according to s)
$\mathrm{sort}(\mathcal{S})$ (Sorts segmentation set $\mathcal{S}$ by variation in DL)
$\mathrm{commit}(\mathcal{S}, \mathbf{\Phi})$ (Apply segmentations in $\mathcal{S}$ to $\mathbf{\Phi}$, returns variation in DL)

```
repeat
    S ← collect(Φ);
    candidates ← [];
    for s ∈ S do
        Δ' ← ΔDL(s, Φ);
        if Δ' ≤ 0 then
            candidates.append({Δ', s});
    sort(candidates);
    Δ ← commit(candidates, Φ);
until Δ > 0;
return Φ;
```

4.1 Description Length Functions

Let us start with the description length $\mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi})$ of a data set $\mathcal{D}$ given a model $\mathbf{\Phi}$. Following information theory [10], the natural approach to compute such description length (in bits) is to use its lower bound given by $\mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi}) = -\log_2(\Pr(\mathcal{D} \mid \mathbf{\Phi}))$.

The description length $\mathrm{DL}(\mathbf{\Phi})$ can be measured as the number of bits required to send model $\mathbf{\Phi}$ over a channel. This can be computed by serializing $\mathbf{\Phi}$ into a sequence of symbols and then computing the length of the optimal encoding of such sequence. To serialize a PB model, we require one symbol for each word in the source and target vocabularies, another symbol to separate the source and target sides in a phrase pair, and one additional symbol to distinguish between the different pairs in the phrase lexicon.

The example toy PB model {La|||The, casa|||house, azul|||blue} will be serialized as "La|The•casa|house•azul|blue", where symbol • separates the phrase pairs, and | separates the two sides of each pair. To compute the description length of the model, we assume a uniform distribution so that each of K different symbols is encoded using $-\log_2(\frac{1}{K})$ bits. In the example, we require 3 bits for each symbol[1], and 33 bits to encode the whole serialized PB model (11 symbols). This uniform code is obviously not optimal. Still, using a common encoding, we can fairly compare different models.

4.2 Inference Procedure

The minimization in Equation (3) requires to a search for an optimal phrase lexicon. Obviously, an exhaustive search over all possible sets of phrase pairs is unfeasible in practice. We follow the ideas in [9] and implement the search as an iterative generalization procedure. Let $\mathcal{D} = \{\mathbf{f}_n, \mathbf{e}_n\}_{n=1}^{N}$ be a data set with N sentence pairs, The initial

[1] We have 8 different symbols: one symbol for each of the six words, plus • and |.

PB model would be a "sentence-based" model that we generalize by identifying parts of the phrase pairs that could be used in isolation. From a probabilistic point of view, this process moves some of the probability mass which is concentrated in the training data out to other data still unseen. Consider this initial "sentence-based" PB model:

La casa azul|||The blue house Esta casa azul|||This blue house
Esta casa verde|||This green house

it can be segmented to obtain a new PB model:

La|||The Esta|||This casa azul|||blue house casa verde|||green house

which explains a new translation (La casa verde→The green house) and has a shorter length (19 symbols vs. 23 original symbols). In [5], only segmentations that bisect the phrases are considered. In Section 5 we propose an extended segmentation approach.

Algorithm 1 describes the MDL PB inference by iterative generalization. First, we collect the potential segmentations of the current PB model (line 2). Then, we estimate the variation in description length due to the application of each segmentation (lines 3 to 7). Finally, we sort the segmentations (line 8) and apply the one with largest length reduction to obtain a new PB model (line 9). The algorithm stops when no reduction is achievable [12, 5]. The number of segmentations under consideration (bisections) is bounded by $O(N \cdot L \cdot M)$, where N is the number of training sentences, and L and M are the maximum length, in words, among the source and target sentences respectively.

4.3 Estimating the Impact of a Segmentation

The key component of Algorithm 1 is function $\Delta\mathrm{DL}(\mathrm{s}, \mathbf{\Phi})$ that evaluates the impact of a candidate segmentation s on the description length of PB model $\mathbf{\Phi}$. That is, $\Delta\mathrm{DL}(\mathrm{s}, \mathbf{\Phi})$ computes the difference in description length between the current model $\mathbf{\Phi}$ and the model $\mathbf{\Phi}'$ that would result from committing to s:

$$\Delta\mathrm{DL}(\mathrm{s}, \mathbf{\Phi}) = \mathrm{DL}(\mathbf{\Phi}') - \mathrm{DL}(\mathbf{\Phi}) + \mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi}') - \mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi}) \tag{4}$$

The length difference between phrase lexicons ($\mathrm{DL}(\mathbf{\Phi}') - \mathrm{DL}(\mathbf{\Phi})$) is given by the difference between the lengths of the phrase pairs added and removed. The difference for the data is given by $-\log_2\left(\frac{\Pr(\mathcal{D}|\mathbf{\Phi}')}{\Pr(\mathcal{D}|\mathbf{\Phi})}\right)$. These probabilities can be computed by translating the training data. However, this is a prohibitively expensive approach. Instead, we estimate the data description length in closed form.

The probability of a phrase pair $\{\tilde{\mathrm{f}}, \tilde{\mathrm{e}}\}$ in a PB model is computed as the number of occurrences of the pair divided by the number of occurrences of the source (or target) phrase [1]. We thus estimate the probabilities in the segmented model $\mathbf{\Phi}'$ by counting the occurrences of the replaced phrase pairs as occurrences of the new segmented pairs. Let $\{\tilde{\mathrm{f}}_0, \tilde{\mathrm{e}}_0\}$ be a phrase pair bisected into $\{\tilde{\mathrm{f}}_1, \tilde{\mathrm{e}}_1\}$ and $\{\tilde{\mathrm{f}}_2, \tilde{\mathrm{e}}_2\}$. The direct phrase probabilities in $\mathbf{\Phi}'$ will be identical to those in $\mathbf{\Phi}$ except that:

$$\mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{e}}_0 \,|\, \tilde{\mathrm{f}}_0) = 0$$

$$\mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{e}}_1 \,|\, \tilde{\mathrm{f}}_1) = \frac{\mathrm{N}_{\mathcal{D}}(\{\tilde{\mathrm{f}}_1, \tilde{\mathrm{e}}_1\}) + \mathrm{N}_{\mathcal{D}}(\{\tilde{\mathrm{f}}_0, \tilde{\mathrm{e}}_0\})}{\mathrm{N}_{\mathcal{D}}(\tilde{\mathrm{f}}_1) + \mathrm{N}_{\mathcal{D}}(\{\tilde{\mathrm{f}}_0, \tilde{\mathrm{e}}_0\})} \quad \mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{e}}_2 \,|\, \tilde{\mathrm{f}}_2) = \frac{\mathrm{N}_{\mathcal{D}}(\{\tilde{\mathrm{f}}_2, \tilde{\mathrm{e}}_2\}) + \mathrm{N}_{\mathcal{D}}(\{\tilde{\mathrm{f}}_0, \tilde{\mathrm{e}}_0\})}{\mathrm{N}_{\mathcal{D}}(\tilde{\mathrm{f}}_2) + \mathrm{N}_{\mathcal{D}}(\{\tilde{\mathrm{f}}_0, \tilde{\mathrm{e}}_0\})}$$

where $\mathrm{N}_{\mathcal{D}}(\cdot)$ are counts in $\mathcal{D}$. Inverse probabilities are computed similarly.

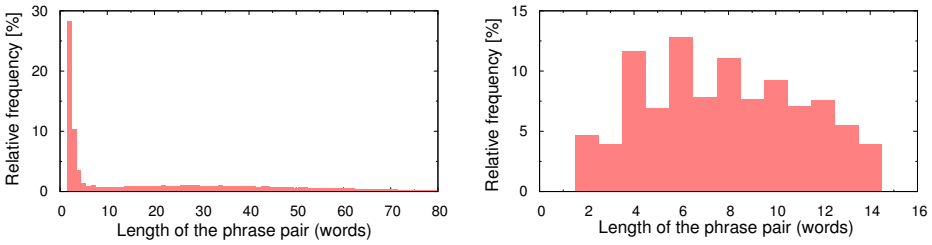

Fig. 1. Histogram of lengths (source plus target words) for two PB models: a conventional MDL model (left) and the same model after further segmenting long phrase pairs (right).

Finally, the variation in data description length is given by the ratio between the estimated probability of the new pairs in $\mathbf{\Phi}'$ and that of the pairs replaced from $\mathbf{\Phi}$:

$$\frac{\Pr(\mathcal{D} \mid \mathbf{\Phi}')}{\Pr(\mathcal{D} \mid \mathbf{\Phi})} \approx \frac{\mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{e}}_1 \mid \tilde{\mathrm{f}}_1) \cdot \mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{e}}_2 \mid \tilde{\mathrm{f}}_2)}{\mathrm{P}_{\mathbf{\Phi}}(\tilde{\mathrm{e}}_0 \mid \tilde{\mathrm{f}}_0)} \cdot \frac{\mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{f}}_1 \mid \tilde{\mathrm{e}}_1) \cdot \mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{f}}_2 \mid \tilde{\mathrm{e}}_2)}{\mathrm{P}_{\mathbf{\Phi}}(\tilde{\mathrm{f}}_0 \mid \tilde{\mathrm{e}}_0)} \tag{5}$$

where the two factors account for the direct and inverse phrase probabilities.

Note that we only consider the direct and inverse phrase probability distributions. A similar approach can be followed to estimate the change in direct and inverse lexical probabilities [2] but we left this extension for future developments.

5 Improving MDL Inference

MDL provides a simple and theoretically-sound approach to perform statistical inference. However, its application to natural language tasks, such as machine translation, presents diverse practical challenges due to the intrinsic sparsity of human language. Frequencies of words in natural language follow a power law [13]. Thus, many of the words (or sequences of words) in a parallel corpus will appear only once. Since MDL "learns" by "finding regularities", this fact imposes a fundamental limitation to the MDL generalization procedure. Consider the following example toy PB model:

La casa azul|||The blue house Este coche verde|||This green car
Tu bicicleta rosa|||Your pink bike

Obviously, there is a clear correspondence between parts of the source and target phrases, e.g. La↔The or coche verde↔green car. However, no bisection provides a reduction in description length. A consequence of this is that MDL tend to estimate phrase lexicons with long phrase pairs, see Figure 1 (left). There, long phrase pairs (more than 15 words[2]) account for almost half the pairs and 91% of the words in the model. This is an important limitation since long phrase pairs generalize poorly.

To address this limitation, we extend the process in Section 4.2 with a subsequent segmentation step to split the remaining long phrase pairs into more general units. Specifically, we implement the state-of-the-art phrase extraction procedure [2]. Figure 1 (right) shows the phrase lengths resulting after such extended segmentation process. Long phrase pairs have been replaced by shorter, more general bilingual phrases.

[2] 14 is the maximum phrase pair length usually considered by conventional PB systems.

	News Commentary (Spa. / Eng.)		
	train	tune	test
#Sentences	51k	2k	1k
#Words	1.4M/1.2M	56k/50k	30k/26k
Vocabulary	47k/35k	5k/5k	8k/7k

Table 1. Main figures of the NC corpus. M and k stand for millions and thousands of elements.

	BLEU [%](↑)	**TER [%](↓)**	**Size**
State-of-the-art	31.4/30.7	48.0/47.2	2.2M
MDL: $\mathrm{DL}(\mathbf{\Phi})$	24.0/23.7	58.3/56.8	79.5k
$+\,\mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi})$	24.0/23.8	57.5/56.0	79.1k
+ Further Segmentation	29.1/28.1	50.6/49.4	1.4M

Table 2. Size (number of phrase pairs) of the inferred MDL PB models, and quality (tune/test) of the generated translations. M and k stand for millions and thousands of elements respectively.

6 Experiments

We conducted experiments on the Spanish-to-English News Commentary (NC) translation task [14], see Table 1. All sentences were tokenized and lowercased.

We inferred MDL PB models with the training partitions as described in Sections 4 and 5. Then, we included them in a log-linear PB model [1] and generated translations for the test partitions [2]. Since we only estimate the direct and inverse phrase probabilities, see Section 4.3, we did not use lexical probabilities [2] in our experiments. Translation quality was measured with BLEU [15] and TER [16]. BLEU measures the accuracy of the automatic translations while TER measures their distance to a reference.

Table 2 shows size (number of phrase pairs) of the inferred MDL PB models, and BLEU and TER scores of their translations. As a comparison, we display results for a state-of-the-art PB system [2]. Results show that the proposed MDL inference obtained much more concise models (less than one tenth the number of phrases) than the standard inference pipeline. However, these smaller models were not able to deliver translations of similar quality which is consistent with previous results reported in [5]. The extended segmentation approach proposed in Section 5 dramatically improved the quality of the generated translations at the cost of an increase in the size of the PB model.

Results obtained considering only the description length of the PB model ($\mathrm{DL}(\mathbf{\Phi})$) and considering the total description length ($\mathrm{DL}(\mathbf{\Phi}) + \mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi})$) were virtually the same. This fact, consistent with previous work [17], indicates that, for PB models, the structure (set of phrase pairs) has the greater impact in translation quality, with only scarce improvements due to the actual probabilities assigned to the phrase pairs.

We also measured the changes of the model and the translation quality during the inference process. Figure 2 displays the number of phrase pairs (left) and the total description length (right) of the intermediate PB models generated during the MDL inference procedure. The total description length is broken down into description length of the model ($\mathrm{DL}(\mathbf{\Phi})$, bottom) and description length of the data given the model ($\mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi})$, top). We can observe that although the number of phrase pairs of the intermediate mod-

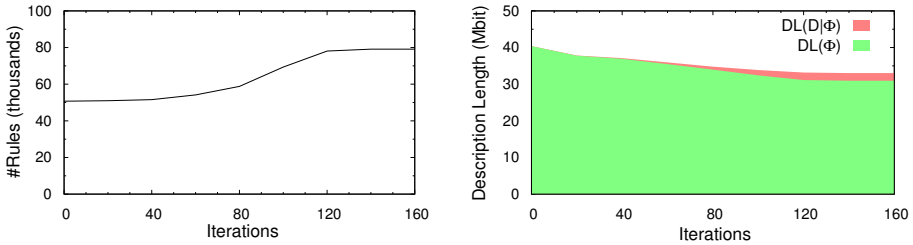

Fig. 2. Changes in the PB model during MDL inference. Left plot displays the number of rules of the model while the plot in the right displays its total description length broken down into model description length (bottom) and data description length given the model (top).

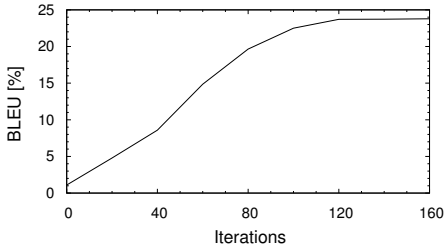

Fig. 3. Variation in test BLEU over iterations.

els increased over time, their description length actually went down. This indicates that we are inferring more general models with a looser fit on the training data. Moreover, the improvements in model description length made up for the loss in data description length which indicates that we were indeed generalizing successfully.

Finally, Figure 3 displays the BLEU for the translations of the test partition over time. Translation quality increased with the iterations reaching its maximum at the end of the inference process. This again came to confirm the soundness of MDL inference.

7 Conclusions and Future Developments

We have studied a simple, unsupervised inference procedure for PB models based on the MDL principle. We have also identified a potential practical limitation of MDL and have proposed an approach to overcome it. Empirical results have shown that the proposed approach was able to boost the quality of MDL PB models.

MDL provides a solid foundation from where to formalize PB inference. Future developments may include (1) a more sophisticated segmentation procedure, (2) the inclusion of lexical models in the MDL inference, and (3) the implementation of a more parsimonious segmentation approach for the long remaining phrase pairs.

Acknowledgments

Work supported by the EU 7^{th} Framework Programme (FP7/2007-2013) under the CasMaCat project (grant agreement n$^{\text{o}}$ 287576), by Spanish MICINN under grant TIN2012-31723, and by the Generalitat Valenciana under grant ALMPR (Prometeo/2009/014).

References

1. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. (2003) 48–54
2. 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., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the Association for Computational Linguistics, demonstration session. (2007)
3. Sanchis-Trilles, G., Ortiz-Martínez, D., González-Rubio, J., González, J., Casacuberta, F.: Bilingual segmentation for phrasetable pruning in statistical machine translation. In: Proceedings of the Conference of the European Association for Machine Translation. (2011)
4. Rissanen, J.: Modeling by shortest data description. Automatica **14**(5) (1978) 465 – 471
5. González-Rubio, J., Casacuberta, F.: Inference of phrase-based translation models via minimum description length. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics. (2014) 90–94
6. Marcu, D., Wong, W.: A phrase-based, joint probability model for statistical machine translation. In: Proceedings of the conference on Empirical methods in natural language processing. (2002) 133–139
7. DeNero, J., Bouchard-Côté, A., Klein, D.: Sampling alignment structure under a bayesian translation model. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2008) 314–323
8. Zhang, J.: Model-Based Search for Statistical Machine Translation. Master's thesis, Edinburgh University, United Kingdom (2005)
9. Vilar, J.M., Vidal, E.: A recursive statistical translation model. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts. (2005) 199–207
10. Shannon, C.: A mathematical theory of communication. Bell System Technical Journal **27** (1948) 379–423, 623–656
11. Grünwald, P.: A tutorial introduction to the minimum description length principle (2004) http://arxiv.org/abs/math/0406077.
12. Saers, M., Addanki, K., Wu, D.: Iterative rule segmentation under minimum description length for unsupervised transduction grammar induction. In: Statistical Language and Speech Processing. Volume LNCS 7978. (2013) 224–235
13. Zipf, G.K.: The Psychobiology of Language. Houghton-Mifflin (1935)
14. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: (Meta-) evaluation of machine translation. In: Proceedings of the Workshop on Statistical Machine Translation. (2007) 136–158
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the Meeting on Association for Computational Linguistics, Association for Computational Linguistics (2002) 311–318
16. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas. (2006) 223–231
17. Turchi, M., De Bie, T., Cristianini, N.: Learning to translate: a statistical and computational analysis. Technical report, University of Bristol (2009)