Inference of Phrase-Based Translation Models via Minimum Description Length

Jesús González-Rubio and **Francisco Casacuberta**
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València, Camino de Vera s/n, 46021 Valencia (Spain)
{jegonzalez, fcn}@dsic.upv.es

Abstract

We present an unsupervised inference procedure for phrase-based translation models based on the minimum description length principle. In comparison to current inference techniques that rely on long pipelines of training heuristics, this procedure represents a theoretically well-founded approach to directly infer phrase lexicons. Empirical results show that the proposed inference procedure has the potential to overcome many of the problems inherent to the current inference approaches for phrase-based models.

1 Introduction

Since their introduction at the beginning of the twenty-first century, phrase-based (PB) translation models (Koehn et al., 2003) have become the state-of-the-art for statistical machine translation (SMT). PB model provide a big leap in translation quality with respect to the previous word-based translation models (Brown et al., 1990; Vogel et al., 1996). However, despite their empirical success, inference procedures for PB models rely on a long pipeline of heuristics (Och and Ney, 2003) and mismatched learning models, such as the long outperformed word-based models. Latter stages of the pipeline cannot recover mistakes or omissions made in earlier stages which forces the individual stages to massively overgenerate hypotheses. This manifests as a huge redundancy in the inferred phrase lexicons, which in turn largely penalizes the efficiency of PB systems at run-time. The fact that PB models usually cannot generate the sentence pairs in which they have been trained in, or that it is even possible to improve the performance of a PB system by discarding most of the learned phrases are clear indicators of these deficiencies (Sanchis-Trilles et al., 2011).

We introduce an unsupervised procedure to infer PB models based on the *minimum description length* (MDL) principle (Solomonoff, 1964; Rissanen, 1978). MDL, formally described in Section 2, is a general inference procedure that "learns" by "finding data regularities". MDL takes its name from the fact that regularities allow to *compress* the data, i.e. to describe it using fewer symbols than those required to describe the data literally. As such, MDL 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 the data and "complexity" or "richness" of the model.

MDL has been previously used to infer monolingual grammars (Grünwald, 1996) and inversion transduction grammars (Saers et al., 2013). Here, we adapt the basic principles described in the latter article to the inference of PB models. The MDL inference procedure, described in Section 3, learns PB models by iteratively generalizing an initial model that perfectly overfits training data. An MDL objective is used to guide this process. MDL inference has the following desirable properties:

- Training and testing are optimized upon the same model; a basic principle of machine learning largely ignored in PB models.
- It provides a joint estimation of the structure (set of bilingual phrases) and the parameters (phrase probabilities) of PB models.
- It automatically protects against overfitting by implementing a trade-off between the expressiveness of the model and training data fitting.

The empirical evaluation described in Section 4 focuses on understanding the behavior of MDL-based PB models and their specific traits. That is, in contrast to a typical PB system building paper, we are not exclusively focused on a short term boost in translation quality. Instead, we aim at studying the adequacy and future potential of MDL as inference procedure for PB models.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 90–94,
Gothenburg, Sweden, April 26-30 2014.

2 The MDL Principle

Given a set of data $\mathcal{D}$, the MDL principle aims at obtaining the simplest possible model $\mathbf{\Phi}$ that describes $\mathcal{D}$ as well as possible (Solomonoff, 1964; Rissanen, 1978). Central to MDL is the one-to-one correspondence between description length functions and probability distributions that follows from the Kraft-McMillan inequality (McMillan, 1956). For any probability distribution $\Pr(\cdot)$, it is possible to construct a coding scheme such that the length (in bits) of the encoded data is minimum and equal to $-\log_2(\Pr(\mathcal{D}))$. In other words, searching for a minimum description length reduces to searching for a good probability distribution, and vice versa. Taking these considerations into account, MDL inference is formalized as:

$$\widehat{\mathbf{\Phi}} = \operatorname*{argmin}_{\mathbf{\Phi}} \mathrm{DL}(\mathbf{\Phi}, \mathcal{D}) \tag{1}$$

$$= \operatorname*{argmin}_{\mathbf{\Phi}} \mathrm{DL}(\mathbf{\Phi}) + \mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi}) \tag{2}$$

where $\mathrm{DL}(\mathbf{\Phi})$ denotes the description length of the model, and $\mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi})$ denotes the description length of the data given the model. A complete introductory tutorial of the MDL principle and methods can be found in (Grünwald, 2004).

3 MDL Phrase-Based Models

3.1 Description Length Functions

We start by defining how to compute $\mathrm{DL}(\mathbf{\Phi})$ and $\mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi})$ for any PB model and data set.

Let $\Pr_{\mathbf{\Phi}}(\mathcal{D})$ be the probability of data set $\mathcal{D}$ according to PB model $\mathbf{\Phi}$. We follow the Kraft-McMillan inequality and define the description length of the data given the model as $\mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi}) = -\log_2(\Pr_{\mathbf{\Phi}}(D))$, which it is the lower bound for the description length of the data.

Regarding the description length of the PB model, $\mathrm{DL}(\mathbf{\Phi})$, we compute it by serializing $\mathbf{\Phi}$ into a sequence of symbols and then computing the length of the optimal encoding of such sequence. To do that, we need one symbol for each word in the source and target languages, 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. For example, the following toy PB model

```
La|||The     casa|||house     azul|||blue
```

is serialized as `La|The•casa|house•azul|blue`, where symbol • separates the phrase pairs, and | separates the two sides of each pair. Assuming a uniform distribution over the K different symbols, each symbol would require $-\log_2(\frac{1}{K})$ bits to encode. We will thus require 3 bits to encode each of the 8 symbols in the example, and 33 bits to encode the whole serialized PB model (11 symbols).

3.2 Inference Procedure

We now describe how to perform the maximization in Equation (2). In the case of PB models, this reduces to a search for the optimal phrase lexicon. Obviously, an exhaustive search over all possible sets of phrase pairs in the data is unfeasible in practice. Following the ideas in (Vilar and Vidal, 2005), we implement a search procedure that iteratively generalizes an initial PB model that perfectly fits the data. Let $\mathcal{D} = \{\mathbf{f}_n, \mathbf{e}_n\}_{n=1}^{N}$ be a data set with N sentence pairs, where $\mathbf{f}_n$ are sentences in the source language and $\mathbf{e}_n$ are their corresponding translation in the target language. Our initial PB model will be as follows:

$$\mathbf{f}_1 \,|||\, \mathbf{e}_1 \quad \cdots \quad \mathbf{f}_n \,|||\, \mathbf{e}_n \quad \cdots \quad \mathbf{f}_N \,|||\, \mathbf{e}_N$$

where the probability of each pair is given by the number of occurrences of the pair in the data divided by the number of occurrences of the source (or target) language sentence.

To generalize this initial PB model, we need to identify parts of the existing phrase pairs that could be validly used in isolation. As a result, the PB model will be able to generate new translations different from the ones in the training data. 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; the very definition of generalization. Consider a PB model such as:

```
   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     casa azul|||blue house
Esta|||This   casa verde|||green house
```

which is able to generate one new sentence pair (`La casa verde`→`The green house`) and has a shorter description length (19 symbols) in comparison to the original model (23 symbols). We only consider segmentations that bisect the source and target phrases. More sophisticated segmentation approaches are beyond the scope of this article.

Algorithm 1 describes the proposed PB inference by iterative generalization. First, we collect the potential segmentations of the current PB

Algorithm 1: Iterative inference procedure.

```
input     : $\mathbf{\Phi}$ (initial PB model)
output    : $\widehat{\mathbf{\Phi}}$ (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)
1  begin
2      repeat
3          $\mathcal{S} \leftarrow \mathrm{collect}(\mathbf{\Phi})$;
4          $\mathrm{candidates} \leftarrow []$;
5          for $\mathrm{s} \in \mathcal{S}$ do
6              $\Delta' \leftarrow \Delta\mathrm{DL}(\mathrm{s}, \mathbf{\Phi})$;
7              if $\Delta' \leq 0$ then
8                  $\mathrm{candidates}\,.\mathrm{append}(\{\Delta', \mathrm{s}\})$;
9          $\mathrm{sort}(\mathrm{candidates})$;
10         $\Delta \leftarrow \mathrm{commit}(\mathrm{candidates}, \mathbf{\Phi})$;
11     until $\Delta > 0$ ;
12     return $\mathbf{\Phi}$;
13 end
```

	EuTransI (Sp / En)		
	train	tune	test
#Sentences	10k	2k	1k
#Words	97k / 99k	23k / 24k	12k / 12k
Vocabulary	687 / 513	510 / 382	571 / 435
OOV	– / –	0 / 0	0 / 0
Perplexity	– / –	8.4 / 3.4	8.1 / 3.3
	News Commentary (Sp / En)		
	train	tune	test
#Sentences	51k	2k	1k
#Words	1.4M / 1.2M	56k / 50k	30k / 26k
Vocabulary	47k / 35k	5k / 5k	8k / 7k
OOV	– / –	390 / 325	832 / 538
Perplexity	– / –	136.2 / 197.9	144.2 / 206.0

Table 1: Main figures of the experimental corpora. M and k stand for millions and thousands of elements respectively. Perplexity was calculated using 5-gram language models.

model (line 3). Then, we estimate the variation in description length due to the application of each segmentation (lines 4 to 8). Finally, we sort the segmentations by variation in description length (line 9) and commit to the best of them (line 10). Specifically, given that different segmentations may modify the same phrase pair, we apply each segmentation only if it only affect phrase pairs unaffected by previous segmentations in $\mathcal{S}$. The algorithm stops when none of the segmentations lead to a reduction in description length. Saers et al., (2013) follow a similar greedy algorithm to generalize inversion transduction grammars.

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:

$$\begin{aligned}\Delta\mathrm{DL}(\mathrm{s}, \mathbf{\Phi}) = {} & \mathrm{DL}(\mathbf{\Phi}') - \mathrm{DL}(\mathbf{\Phi}) \\ & + \mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi}') - \mathrm{DL}(\mathcal{D} \mid \mathbf{\Phi}) \quad (3)\end{aligned}$$

The length difference between the phrase lexicons $(\mathrm{DL}(\mathbf{\Phi}') - \mathrm{DL}(\mathbf{\Phi}))$ is trivial. We merely have to compute the difference between the lengths of the phrase pairs added and removed. The difference for the data is given by $-\log_2\left(\frac{\mathrm{Pr}_{\mathbf{\Phi}'}(\mathcal{D})}{\mathrm{Pr}_{\mathbf{\Phi}}(\mathcal{D})}\right)$, where $\mathrm{Pr}_{\mathbf{\Phi}'}(\mathcal{D})$ and $\mathrm{Pr}_{\mathbf{\Phi}}(\mathcal{D})$ are the probability of $\mathcal{D}$ according to $\mathbf{\Phi}'$ and $\mathbf{\Phi}$ respectively. These probabilities can be computed by translating the training data. However, this is a very expensive process that we cannot afford to perform for each candidate segmentation. Instead, we estimate the description length of the data in closed form based on the probabilities of the phrase pairs involved. The probability of a phrase pair $\{\tilde{\mathrm{f}}, \tilde{\mathrm{e}}\}$ is computed as the the number of occurrences of the pair divided by the number of occurrences of the source (or target) phrase. We thus estimate the probabilities in the segmented model $\mathbf{\Phi}'$ by counting the occurrences of the replaced phrase pairs as occurrences of the segmented pairs. Let $\{\tilde{\mathrm{f}}_0, \tilde{\mathrm{e}}_0\}$ be the phrase pair we are splitting 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:

$$\begin{aligned}\mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{e}}_0 \mid \tilde{\mathrm{f}}_0) &= 0 \\ \mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{e}}_1 \mid \tilde{\mathrm{f}}_1) &= \frac{\mathrm{N}_{\mathbf{\Phi}}(\{\tilde{\mathrm{f}}_1, \tilde{\mathrm{e}}_1\}) + \mathrm{N}_{\mathbf{\Phi}}(\{\tilde{\mathrm{f}}_0, \tilde{\mathrm{e}}_0\})}{\mathrm{N}_{\mathbf{\Phi}}(\tilde{\mathrm{f}}_1) + \mathrm{N}_{\mathbf{\Phi}}(\{\tilde{\mathrm{f}}_0, \tilde{\mathrm{e}}_0\})} \\ \mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{e}}_2 \mid \tilde{\mathrm{f}}_2) &= \frac{\mathrm{N}_{\mathbf{\Phi}}(\{\tilde{\mathrm{f}}_2, \tilde{\mathrm{e}}_2\}) + \mathrm{N}_{\mathbf{\Phi}}(\{\tilde{\mathrm{f}}_0, \tilde{\mathrm{e}}_0\})}{\mathrm{N}_{\mathbf{\Phi}}(\tilde{\mathrm{f}}_2) + \mathrm{N}_{\mathbf{\Phi}}(\{\tilde{\mathrm{f}}_0, \tilde{\mathrm{e}}_0\})}\end{aligned}$$

where $\mathrm{N}_{\mathbf{\Phi}}(\cdot)$ are counts in $\mathbf{\Phi}$. Inverse probabilities are computed accordingly. Finally, we compute the variation in data description length using:

$$\begin{aligned}\frac{\mathrm{Pr}_{\mathbf{\Phi}'}(\mathcal{D})}{\mathrm{Pr}_{\mathbf{\Phi}}(\mathcal{D})} \approx {} & \frac{\mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{e}}_1 \mid \tilde{\mathrm{f}}_1) \cdot \mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{e}}_2 \mid \tilde{\mathrm{f}}_2)}{\mathrm{P}_{\mathbf{\Phi}}(\tilde{\mathrm{e}}_0 \mid \tilde{\mathrm{f}}_0)} \\ & \cdot \frac{\mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{f}}_1 \mid \tilde{\mathrm{e}}_1) \cdot \mathrm{P}_{\mathbf{\Phi}'}(\tilde{\mathrm{f}}_2 \mid \tilde{\mathrm{e}}_2)}{\mathrm{P}_{\mathbf{\Phi}}(\tilde{\mathrm{f}}_0 \mid \tilde{\mathrm{e}}_0)} \quad (4)\end{aligned}$$

	EUtransI		News Commentary	
	BLEU [%] (tune/test)	Size	BLEU [%] (tune/test)	Size
SotA	91.6 / 90.9	39.1k	31.4 / 30.7	2.2M
MDL	88.7 / 88.0	2.7k	24.8 / 24.6	79.1k

Table 2: Size (number of phrase pairs) of the MDL-based PB models, and quality of the generated translations. We compare against a state-of-the-art PB inference pipeline (SotA).

For a segmentation set, we first estimate the new model Φ' to reflect all the applied segmentations, and then sum the differences in description length.

4 Empirical Results

We evaluated the proposed inference procedure on the EuTransI (Amengual et al., 2000) and the News Commentary (Callison-Burch et al., 2007) corpora. Table 1 shows their main figures.

We inferred PB models (set of phrase pairs and their corresponding probabilities) with the training partitions as described in Section 3.2. Then, we included these MDL-based PB models in a conventional log-linear model optimized with the tuning partitions (Och, 2003). Finally, we generated translations for the test partitions using a conventional PB decoder (Koehn et al., 2007).

Table 2 shows size (number of phrase pairs) of the inferred MDL-based PB models, and BLEU score (Papineni et al., 2002) of their translations of the tune and test partitions. As a comparison, we display results for a state-of-the-art (SotA) PB system (Koehn et al., 2007). These results show that MDL inference obtained much more concise models (less than one tenth the number of phrases) than the standard inference pipeline. Additionally, the translations of the simple EuTransI corpus were of a similar quality as the ones obtained by the SotA system. In contrast, the quality of the translations for News Commentary was significantly lower.

To better understand these results, Figure 1 displays the histogram of phrase lengths (number of source words plus target words) of the SotA model and the MDL-based model for the News Commentaries corpus. We first observed that the length of the phrase pairs followed a completely different distribution depending on the inference procedure. Most of the phrase pairs of the MDL-based model translated one source word by one target word with an exponential decay in frequency for longer phrase pairs; a typical distribution of events in natural language (Zipf, 1935). Longer phrase pairs, about 45% of the total, contain sequences of words that only appear once in the corpus, and thus, they cannot be segmented in any way that leads to a reduction in description length. Although formally correct, long phrase pairs generalize poorly which explains the comparatively poor performance of MDL inference for the News Commentaries corpus. This problem was largely attenuated for EuTransI due to its simplicity.

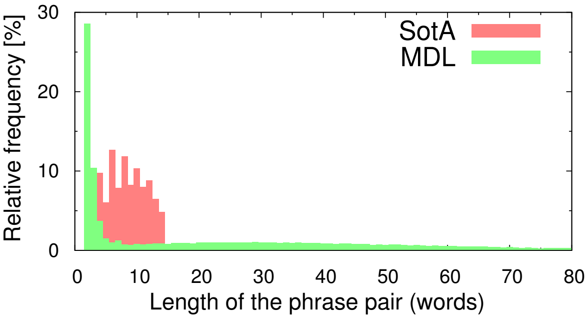

Figure 1: Histogram of lengths (source plus target words) for the phrase pairs in the inferred models.

5 Conclusions and Future Developments

We have described a simple, unsupervised inference procedure for PB models that learns phrase lexicons by iteratively splitting existing phrases into smaller phrase pairs using a theoretically well-founded minimum description length objective. Empirical results have shown that the inferred PB models, far from the artificial redundancy of the conventional PB inference pipeline, are very parsimonious and provide competitive translations for simple translation tasks.

The proposed methodology provides a solid foundation from where to develop new PB inference approaches that overcome the problems inherent to the long pipeline of heuristics that nowadays constitute the state-of-the-art. Future developments in this direction will include:

- A more sophisticated segmentation procedure that allow to divide the phrases into more that two segments.
- A hybrid approach where the long phrase pairs remaining after the MDL inference are further segmented, e.g., according to a word lexicon.
- The inclusion of lexical models in the definition of the PB model.

Acknowledgments

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

References

Juan-Carlos Amengual, M. Asunción Castaño, Antonio Castellanos, Víctor M. Jiménez, David Llorens, Andrés Marzal, Federico Prat, Juan Miguel Vilar, José-Miguel Benedí, Francisco Casacuberta, Moisés Pastor, and Enrique Vidal. 2000. The eutrans spoken language translation system. *Machine Translation*, 15(1-2):75–103.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 136–158.

Peter Grünwald. 1996. A minimum description length approach to grammar inference. *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pages 203–216.

Peter Grünwald. 2004. A tutorial introduction to the minimum description length principle. http://arxiv.org/abs/math/0406077.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics, demonstration session*, June.

Brockway McMillan. 1956. Two inequalities implied by unique decipherability. *IRE Transactions on Information Theory*, 2(4):115–116.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Meeting on Association for Computational Linguistics*, pages 160–167. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica*, 14(5):465 – 471.

Markus Saers, Karteek Addanki, and Dekai Wu. 2013. Iterative rule segmentation under minimum description length for unsupervised transduction grammar induction. In *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 224–235. Springer.

Germán Sanchis-Trilles, Daniel Ortiz-Martínez, Jesús González-Rubio, Jorge González, and Francisco Casacuberta. 2011. Bilingual segmentation for phrasetable pruning in statistical machine translation. In *Proceedings of the 15th Conference of the European Association for Machine Translation*.

Ray Solomonoff. 1964. A formal theory of inductive inference, parts 1 and 2. *Information and Control*, 7:1–22, 224–254.

Juan Miguel Vilar and Enrique Vidal. 2005. A recursive statistical translation model. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 199–207.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, pages 836–841.

George Kingsley Zipf. 1935. *The Psychobiology of Language*. Houghton-Mifflin.