Introduction to Probabilistic Models in Natural Language Generation

Leo Leppänen
leo.leppanen@helsinki.fi

I. Introduction

Generative probabilistic models are a diverse category of different ways of modeling a process that produces output from input. Common to all of them is that they are probabilistic; in a sense, they model the processes as consisting of decision steps, where at each step the next action is chosen randomly from a set of possible choice, where each alternative has a distinct probability of being chosen.

In this text, we will focus on two specific probabilistic models that attempt to mimic the way a human generates natural language reports of numerical or categorical data. The focus will be largely on the intuitions of these models, and less on the rigorous mathematics of them. This decision to disregard the mathematics is caused by two things. First, as this text is intended for an audience with a varied mathematical background, including the necessary basics of probability theory would make this text unnecessarily long. Second, the models themselves are actually fairly simple, insofar as probabilistic models are considered; there is very little value in describing them to the mathematically inclined.

II. Writing as a Process

The two models we are discussing are the model proposed by Liang et al. [1] and the model proposed by Angeli et al. [2]. They both consider the process of writing as consisting of three larger stages: *record selection*, *field-set selection* and *representation selection*. The process takes as input *knowledge* and outputs a text.

The knowledge is assumed to consist of *records*, where each record is somewhat analogous to a row in a database. The records have a type that defines what kind of records they are: one type of record might contain the information about the minimum, maximum and mean temperature of a single day, whereas another record might contain the total rainfall of that same day. [1], [2]

These records consist of *fields*, where again each field has a type and a value. The minimum, maximum and mean temperatures of the day would each be a distinct field with their distinct values. The fields also have a type that defines what type of content they can contain: numeric, textual or categorical. These field-types and the number of fields within a record are defined by the record type and are the same across all records of the same type. Only the values of the fields vary. [1], [2]

The first stage of the three stage process corresponds to choosing the records that the text should discuss. In the second stage some fields from those records are chosen. Finally, in the third stage the surface representations of those fields are selected.

III. The Liang Model

Liang et al. [1] assume that the stages are completed one after another: first, all the record selections are made consecutively. Second, for all those records one or more fields are chosen. Finally, for each field in turn a sequence of words is selected one by one.

As the model is probabilistic, it assumes that for each selection there are a multitude of options, each with a distinct probability of being chosen as the one to take. These probabilities are affected by certain factors. When a record is being selected, the probabilities of all the possible records are affected by both *coherence* and *salience*: whether this record is likely to follow the previously chosen record, and whether this record is likely to be chosen at all. When selecting the fields for a record, the probability of a field is dependent on the previous field selected for that same record: weather reports often list either the minimum and the maximum temperature, or only the mean temperature. The final third stage first chooses a number of words to pick – based on an uniform distribution from zero a pre-set maximum – and then selects that number of words from a distribution specific to the record the field. [1]

Some extra complications are taken with respect to the fields: for example for each numerical value in the original data, the model considers six distinct values that it could pick: the original value, the floor and the ceil of the original value, a rounded value and the original value modified either up or down by a small amount [1]. This reflects the fact that a weather report could say “the temperature averaged around 25”, even if the actual mean temperature was for example 25.9 or 27 degrees.

The model is learned from a set of training data, where for each training sample both the inputs and the outputs are known. Technically speaking, the actual training uses an expectation maximization algorithm to maximize the marginal distribution of the training data, repeating an E-step of computing expected counts according to the posterior distribution of the latent variables given the training data and the current parameters, and an M-step of optimizing the parameters based on the expected counts. In more intuitive terms, the training data is analyzed and the internal probabilities of the decision the model makes are set so that the probabilities of the training outputs are as high as possible when the inputs are the training inputs. [1]

While the model could be used for generation of text as well, Liang et al. [1] actually only evaluate the method as a tool for semantic alignment. Still, the it is important to understand as it is a critical part of the next model.

IV. The Angeli Model

The Angeli et al. [2] model models the writing process as three distinct types of phases, similar to the Liang model. The largest difference between the models is in how they structure these stages. Recall that Liang et al. ordered the stages so that first all records were selected, followed by the selection of all fields, followed by the selection of all the words. The first modification Angeli et al. make is that they replace the word-by-word selection process by a selection of a single *template* for each set of fields. Second, instead of making all the selections of a single stage consecutively, they mix the stages up so that first one record is selected, then the fields for that record are selected, then the template for those fields are selected, and finally the process is repeated.

Furthermore, they let a larger amount of features influence the decision processes at each decision point. For record selection, they consider coherence both on a local level (Is this record likely to be chosen given the previous record?) and on a global level (Is this record likely to be chosen given all the previously chosen records?), whether a record is likely to be chosen if a record of same type was also previous chosen and whether the record's value is such that the record is likely to be chosen. For fields, the tendency of the fields to appear together is considered, as are the fields' values. Finally, the choice of template is affected by how common that template is in general, the values of the fields that go into the template and the language model. [2]

When learning the model, Angeli et al. [2] actually depend heavily on the model of Liang et al. [1] by using it to extract the templates from the training samples. After that, they train their model in a similar fashion to the Liang et al. method. A somewhat large technical difference is that instead of using an EM-algorithm, they simply use a powerful optimization algorithm to maximize the likelihood of the training data. On the level of intuition, this is essentially equivalent to what Liang et al. do, just with different tools.

Once learning is complete, the model is used to generate text by giving it a new input, and then by following the model's process, making all choices randomly from the distributions learned during the training process. This same method would be used with the Liang et al. model as well, if one were to generate text with it.

Angeli et al. [2] used their model to generate texts of three different types: marine wind forecasts, general weather forecasts and robot football play-by-play commentaries. It is notable that all of these text types are very short, with the play-by-plays being only 5.7 words long on average. The longest texts were the general weather forecasts, averaging 28.7 words, but in their case a huge amount of 29,528 training samples was used. This raises doubts about whether these methods would be able to perform well with longer texts.

Ignoring these doubts on the generalizability, Angeli et al. report fairly good results. Evaluated on a 5-point scale on both semantic correctness and fluency (5 being the highest, namely "flawless" and "perfect"), variants of their models scored essentially equivalently to humans controls in fluency in all three text contexts. For semantic correctness, the Angeli models averaged worse scores than human controls in two cases and fared slightly better in the third (general weather reports). But even in the case of the semantic correctness, in all three contexts the worst human controls were rated lower than the best machine produced texts. For the most complex case – general weather reports – Angeli et al. reports a BLEU-score [3] of 51.5. [2]

V. Conclusions

Probabilistic models such as those presented by Liang et al. [1] and Angeli et al. [2] are intriguing, in that they approximate the very complex human behaviors using very simplistic models that still manage to perform surprisingly well when compared to the results of the humans.

It is notable that both of the models presented herein are completely domain independent: they make absolutely no assumptions about what the data is and what it means. Perhaps more interestingly, they are largely language independent; they are likely to perform equivalently well with languages that inflect as little as English. How they do with more synthetic languages (such as Finnish) is more uncertain, but the outlook is not good; at the minimum, more training samples is required for the models to learn all the inflections.

At the same time, the methods suffer from the same problems that all other methods that depend on training data suffer from; the output of a model is ever only as good as the data that is used to build the model. If the aim of the system is to produce large amounts of variety in longer texts, the required amount of training data can become simply impossible to acquire in most contexts. For example, it is questionable whether the total amount of election news written by the Finnish media total the nearly thirty thousand samples that Angeli et al. used in training their weather reporting model, a context of significantly less complexity.

References

[1] P. Liang, M. I. Jordan, and D. Klein, "Learning semantic correspondences with less supervision," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 91–99.

[2] G. Angeli, P. Liang, and D. Klein, "A simple domain-independent probabilistic approach to generation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 502–512.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.