Statistical Structure in Language Processing
IBM Model 1

Cristina Gârbacea
10407936
cr1st1na.garbacea@gmail.com

Sara Veldhoen
10545298
sara.veldhoen@student.uva.nl

Abstract

In this paper we present the theoretical notions underlying the IBM Model 1 we implemented and the Expectation Maximization algorithm we applied to train the given parallel Dutch - English corpus. We show that this model can easily overfit and that maximizing the likelihood of the training data can result in low performance accuracy. For this reason we came with our own extension to the model, which we compare and evaluate against the basic IBM model.

1 Introduction

Statistical alignment models are widely used nowadays in a variety of natural language processing applications, ranging from machine translation to question answering and information retrieval. The goal of finding the best English translation $\mathbf{e}$ of a foreign sentence $\mathbf{f}$ is modeled under the assumption of the *noisy channel* hypothesis, which considers the foreign sentence as a "corrupted" instance of the original English sentence. The English sentence which is at the source of its foreign counterpart is an unknown issue, and thus the task of translating $\mathbf{f}$ into $\mathbf{e}$ becomes one of maximizing the probability of the English sentence given the foreign sentence, $P(\mathbf{e}|\mathbf{f})$. According to Bayes' theorem,

$$P(\mathbf{e}|\mathbf{f}) = \frac{P(\mathbf{e}) \cdot P(\mathbf{f}|\mathbf{e})}{P(\mathbf{f})} \qquad (1)$$

the translation problem can be expressed as:

$$P(\mathbf{e}|\mathbf{f}) = arg \max_{\mathbf{e}} P(\mathbf{e}) \cdot P(\mathbf{f}|\mathbf{e}) \qquad (2)$$

where $P(\mathbf{e})$ denotes the language model, that takes care of the fluency of the output, and $P(\mathbf{e}|\mathbf{f})$ denotes the translation model, that makes sure the translation is adequate.

In order to be able to approximate this probability, we need to make assumptions. In a series of increasingly complex models is presented. We implemented the first model, hereafter IBM model 1. In what follows we present the this model in section 2, together with the Expectation Maximization formula used for training this model. We came up with an improvement over this model, which we introduce in section 4. Then comes an evaluation of the alignment quality for both models in section 5 and finally we conclude in Section 6.

2 IBM Model 1

The assumption underlying the IBM models is that translating text comes down to aligning the words of the foreign and the English sentence, and translating the words independently. The words of the English sentence are called *cepts* and generally within the IBM models *groups* of cepts cannot be aligned to *groups* of words in the foreign sentence. Rather, each cept is aligned to one foreign word or, in case there is no foreign correspondent, to the so called *null* word. Because of this simplification, an alignment can be represented as a vector that has the same length as the English sentence.

Hence the translation probability $P(\mathbf{f}|\mathbf{e})$ can be rewritten as the sum over all possible alignments $\mathbf{a}$ of conditional probabilities $P(\mathbf{f}, \mathbf{a}|\mathbf{e})$:

$$P(\mathbf{f}|\mathbf{e}) = \sum_a P(\mathbf{f}, \mathbf{a}|\mathbf{e}) \qquad (3)$$

The conditional probability $P(\mathbf{f}, \mathbf{a}|\mathbf{e})$ can in turn be expressed as:

$$P(m|\mathbf{e}) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) P(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e}) \qquad (4)$$

The translation probability of a foreign sentence $f = (f_1, \ldots, f_i \ldots, f_{l_f})$ into an English sentence

$e = (e_1, \ldots, e_j, \ldots, e_{l_e})$ based on an alignment function $a : j \rightarrow i$ is defined as:

$$p(e, a|f) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j | f(a_j)) \quad (5)$$

3   EM Training Formula

The parameters of IBM Model 1 for a given pair of languages are estimated using the EM algorithm. It takes as training data a corpus of paired sentences, each pair made up of a sentence in one language and its translation in the other language. Each target word is assumed to be generated by exactly one source word, including the *null* word. The initial step of EM consists in setting uniform translation probabilities $t(e|f)$ over the target vocabulary. During the E-step the model is applied to the data and the alignment probabilities are computed; during the M-step counts over all possible alignments are collected and weighted by their probability.
   + training formula..

4   Improvement Over IBM Model 1

5   Experiments and Results

6   Conclusion

References

[1] Philipp Koehn 2010. *Statistical Machine Translation*. Cambridge University Press.