Statistical Structure in Language Processing
IBM Model 1

Cristina Gârbacea
10407936
cr1st1na.garbacea@gmail.com

Sara Veldhoen
sara.veldhoen@gmail.com

Abstract

Given a parallel Dutch - English corpus, 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 corpus. We are trying to 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 e of a foreign sentence 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 f into e becomes one of maximizing the probability of the English sentence given the foreign sentence, $P(e|f)$. According to Bayes' theorem,

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

the translation problem can be expressed as:

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

where $P(e)$ denotes the language task and $P(e|f)$ denotes the translation task.

IBM models focus on the translation task only which assumes that word alignments exists between the words of the foreign and the English sentence. The words of the English sentence are called *cepts* and generally within the IBM models groups of cepts cannot be aligned to group of words from the other sentence. The rule is that each foreign word or group of foreign words can be aligned to at most one English word. To counter against the case when English words have no foreign correspondent, the *null* word is added as well to the foreign vocabulary, resulting in a fully defined alignment function. Hence the translation probability $P(f|e)$ can be rewritten as the sum on all alignments of conditional probabilities $P(f, a|e)$:

$$P(f|e) = \sum_{a} P(f, a|e) \tag{3}$$

where a represents one possible alignment between the foreign and the English sentences. The conditional probability $P(f, a|e)$ can in turn be expressed as:

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

In what follows we present the IBM Model in Section 2, the Expectation Maximization formula used for training this model in Section 3, our improvements over this model in Section 4, evaluation of the alignment quality in Section 5 and finally we conclude in Section 6.

2 IBM Model 1

IBM models build upon one another in increasing order of complexity. IBM Model 1 is the simplest probabilistic generative model based on lexical translation which assumes a word-to-word

mapping between the target and the source sentence. It is widely used in working with parallel bilingual corpora, aligning syntactic fragments and estimating phrase translation probabilities.

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:

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

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.

4 Improvement Over IBM Model 1

5 Experiments and Results

6 Conclusion

References

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