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 also came up with our own extension to the model, which we compare and evaluate against the basic IBM model 1.

1 Introduction

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,

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

the translation problem can be expressed as:

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

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 (1), (2) a series of increasingly complex models is presented. We implemented the first model, hereafter IBM model 1.

In what follows we present 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 3. Then comes an evaluation of the alignment quality for both models in section 4 and finally we conclude in Section 5.

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 $\mathbf{e}$ are called *cepts* and generally within the IBM models groups of words in the foreign sentence $\mathbf{f}$ cannot be aligned to groups of cepts. Rather, each foreign word is aligned to one cept or, in case there is no English correspondent, to the so called *null* word. Because of this simplification, an alignment can be represented as a vector $\mathbf{a}$ that has the same length as the foreign sentence: m.

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

$$\begin{aligned} P(\mathbf{f}|\mathbf{e}) &= \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e}) \\ &= \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}, m|\mathbf{e}) \\ &= \sum_{\mathbf{a}} P(m|\mathbf{e}) \times P(\mathbf{f}|\mathbf{a}, m, \mathbf{e}) \\ &= P(m|\mathbf{e}) \sum_{\mathbf{a}} P(\mathbf{f}|\mathbf{a}, m, \mathbf{e}) \end{aligned} \tag{3}$$

$$\begin{aligned} P(\mathbf{f}|\mathbf{a}, m, \mathbf{e}) = &\prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \\ &\times P(f_j|a_1^{j}, f_1^{j-1}, m, \mathbf{e}) \end{aligned} \tag{4}$$

EM Training The input to the training is a corpus of paired sentences. To estimate the latent variables, alignments, a version of expectation maximization is used. Training focuses on the pa-

rameters in equation 4. The first term, the alignment probability, is assumed to be uniform in IBM Model 1.

The initial step of EM consists in setting uniform translation probabilities $t(e|f)$ over the target vocabulary.

In the E-step, the alignment probabilities are computed. Note that this is done implicitly, in the process of gathering the fractional counts:

$$c(f|e;\mathbf{f},\mathbf{e}) = \frac{t(f|e)}{t(f|e_0)+\ldots+t(f|e_l)} \times \sum_{j=1}^{m}\delta(f,f_j)\sum_{i=0}^{l}\delta(e,e_i) \quad (5)$$

The translation probabilities are re-estimated in the M-Step, using the obtained counts:

$$t(f|e) = \lambda_e^{-1}\sum_{s=1}^{S} c(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)}) \quad (6)$$

$$\lambda_e = \sum_{f}\sum_{s=1}^{S} c(f|e;\mathbf{f}^{(s)},\mathbf{e}^{(s)}) \quad (7)$$

Where equation 7 (λ_e) defines the normalization parameter to get proper probabilities.

3 Improvement Over IBM Model 1

In IBM Model 1, all alignments are equally probable. The Viterbi alignment can thus be found by simply choosing, for each word, the cept that has the highest translation probability.

An improvement can be made by introducing a probability function over alignments: $p(a|m,l)$. For simplification, we assume independence of each alignment decision, such that we can compute $p(a_j|j,m,l)$ for each word.

This is also done in IBM Model 2. However, instead of training the alignment probabilities, as in IBM Model 2, we use a heuristic function. We base this function on the assumption that a word in the input should stay close, i.e. we want to favor $a_j = j$ over $a_j > j$ and $a_j < j$. Furthermore, we tried to favor choosing the null word as presented in (3), but that resulted in a drop in performance.

4 Experiments and Results

We implemented IBM Model 1 and trained it with 20 iterations of the Expectation Maximization algorithm on the given Dutch-English parallel corpus consisting of 1.000 sentences. We evaluated the alignment quality against the quality of the alignments provided by Giza++. Below we report the results we obtained, as well as the results of our approach:

	Precison	**Recall**	**F1-score**
IBM Model 1	0.9111	0.9053	0.9063
Our extension	0.7995	0.8901	0.8378

Table 1: Results after training IBM Model 1 with 20 Expectation Maximization iterations

We notice a drop in precision, while the recall tends to be around the same value. Looking at Figure 1 we infer that our extension tends to work better for short sentences than for longer sentences as compared to the original IBM Model 1.

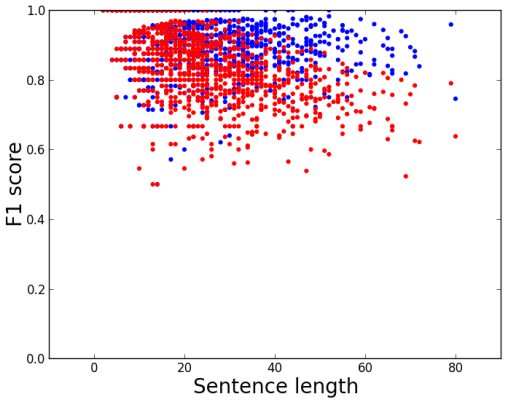

Figure 1: Scatter plot of F1 scores against sentence length. The blue dots denote the original IBM Model 1, while the red dots denote our extension.

5 Conclusion

References

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

[2] Brown, Peter F and Pietra, Vincent J Della and Pietra, Stephen A Della and Mercer, Robert L 1993. *The mathematics of statistical machine translation: Parameter estimation*. Computational linguistics. 19.2: 263-311.

[3] Moore, Robert C. 2004 *Improving IBM word-alignment model 1*. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics.