Statistical Machine Translation

Sophie Arnoult, Gideon Maillette de Buy Wenniger and Andrea Schuch

December 7, 2010

1 Introduction

All the IBM models, and Statistical Machine Translation (SMT) in general, model the problem of finding the best English translation for a French sentence[1] as a noisy channel: The French sentence $\mathbf{f}$[2] to translate is considered to be a 'corruption' of the English sentence $\mathbf{e}$ as depicted in Figure 1. Which English sentence is at the source of its French counterpart is unknown, but we can look for sentences that maximise $P(\mathbf{e}|\mathbf{f})$.

Figure 1: The noisy channel model of machine translation

As by Bayes' theorem,

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

the best English translation for a French sentence is:

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

The translation problem is thus reduced to a language task, to obtain $P(\mathbf{e})$,

[1] We follow the convention that the source language of the translation task is French, and the target language English.

[2] See the Appendix for a list of notations

and a translation task, to obtain $P(\mathbf{f}|\mathbf{e})$[3].

The IBM models focus on the translation task only. To emphasize that the English sentence represents a 'collection' of concepts, the words of the sentence are called *cepts*. Each cept generates one or more words in the French translation. Conversely, each word in the French sentence is said to be *aligned* to a cept in the English sentence. To account for the possibility that French words have no counterpart in the English translation, one says that these words are aligned to the *empty cept*, e_0. Figure 2 shows a possible alignment between two sentences. In this case, the last three words of the French sentence are aligned to the same word in the English sentence. Note that the IBM model of alignment does not allow for the reverse, that is to say, a French word cannot be aligned with more than one English word. Generally, a group of words cannot be aligned to another group of words with the IBM models. This would in fact require a Phrase-Based approach.

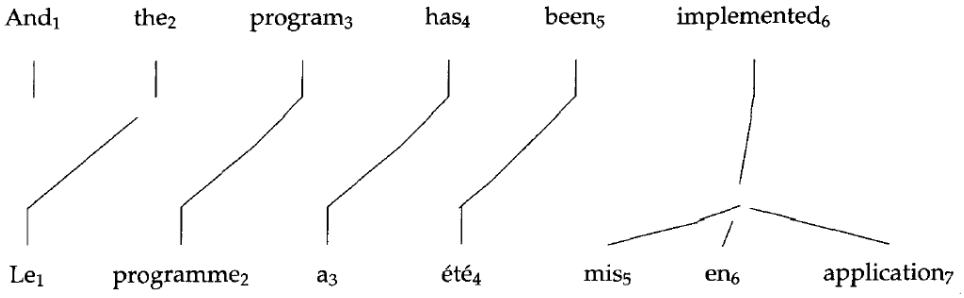

Figure 2: Alignment between a French and an English sentence, after [1]. In the IBM notation, this alignment is represented as [2, 3, 4, 5, 6, 6, 6].

Each IBM model builds on the previous one, each increasing the complexity of the alignment model. In all cases, the translation probability $P(\mathbf{f}|\mathbf{e})$ is seen as the sum on all alignments of the conditional probabilities $P(\mathbf{f}, \mathbf{a}|\mathbf{e})$, where $\mathbf{a}$ is an alignment between the French and the English sentences:

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

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

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

[3]The translation model informs us on what sentences are good translations, while the language model ensures that these sentences are well-formed. By combining these models, we thus get better results than if we were to look directly for the sentence that maximises $P(e|f)$.

The following sections present the main assumptions taken by models 1 to 3, along with the principal equations and the algorithm proposed to compute $P(\mathbf{f}|\mathbf{e})$.

1.1 Model 1

Model 1 takes $P(m|\mathbf{e})$ to be a parameter ϵ independent of $\mathbf{e}$ and m. The term $P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e})$ is assumed to depend only on l and to be equal to $(l+1)^{-1}$, and $P(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e})$ is assumed to depend only on f_j and e_{a_j}, and can thus be expressed as the conditional probability $t(f_j|e_{a_j})$. Equation 4 can therefore be rewritten as:

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{5}$$

Importantly, these assumptions make it possible to compute $P(\mathbf{f}|\mathbf{e})$ efficiently, as Equation 3 can be rewritten as:

$$P(\mathbf{f}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i) \tag{6}$$

The translation probabilities $t(f_j|e_i)$ can be computed using:

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

where:

$$c(f|e; \mathbf{f}, \mathbf{e}) = \frac{t(f|e)}{t(f|e_0) + ... + t(f|e_l)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=0}^{l} \delta(e, e_i) \tag{8}$$

and λ_e is a normalisation parameter ($\sum_f t(f|e) = 1$):

$$\lambda_e = \sum_f \sum_{s=1}^{S} c(f|e; \mathbf{f}^{(s)}, \mathbf{e}^{(s)}) \tag{9}$$

The algorithm used to compute the translation probabilities $t(f_j|e_i)$ is the following:

1. Choose initial values for $t(f|e)$. As $P(\mathbf{f}|\mathbf{e})$ has a unique local maximum in this model, it is sufficient to take these initial values to be equal and different from zero.

2. Compute the counts $c(f|e; \mathbf{f}^{(S)}, \mathbf{e}^{(S)})$ for each pair of sentences using Equation 8.

3. For each e appearing in the English sentences, compute λ_e using Equation 9, and $t(f|e)$ using Equation 7.

4. Repeat steps 2 and 3 until the values of $t(f|e)$ have converged to the desired degree.

A Notations

e	English word
f	French word
e	English sentence
f	French sentence
l	length of English sentence
m	length of French sentence
i	word index in English sentence
j	word index in French sentence
e_0	the empty cept
a	alignment between a French and an English sentence
a_j	the index of the English word that is aligned to the French word at index j

References

[1] Brown et. Al. (1993): The mathematics of statistical machine translation: parameter estimation. Computational Linguistics Vol. 19, No. 2

[2] The Europarl parallel corpus: http://www.statmt.org/europarl/