Alignment models for statistical translation

Thomas Debarre, Matthieu Kirchmeyer, Sylvain Truong, Nicolas Zhang

Introduction

This project explores three methods to solve the problem of word alignments for a bilingual corpus from the conventional mixtures models (IBM1, IBM2) to a first-order HMM model developped in [2].

The goal is to translate a text given in a language (French in all that follows) to another language (English) taking into account one-to-many alignments or null words; the quality of the translation relies on the quality of the word alignments.

We evaluated our algorithms on a small bilingual French-English dataset $\mathcal{D} = \{(\mathbf{f}^{(1)}, \mathbf{e}^{(1)}), \ldots, (\mathbf{f}^{(N)}, \mathbf{e}^{(N)})\}$ which consists of $N = 22$ French sentences composed of 4 to 10 words with their translated English equivalents.

Statistical translation models

IBM models

The IBM alignment models for statistical machine translation were introduced by IBM more than 20 years ago [1]. In this project, we implemented the 2 most basic IBM models, IBM1 and IBM2. Both these models are based on the decomposition of the joint probability of a French sentence $\mathbf{f}$ of length I and its alignment $\mathbf{a}$ conditioned on the English sentence $\mathbf{e}$ of length J:

$$p(\mathbf{f}, \mathbf{a}|\mathbf{e}) = p(I|J) \prod_{i=1}^{I} p(a_i|i, J, I) \cdot p(f_i|e_{a_i})$$

The alignment vector $\mathbf{a} = (a_1, \ldots, a_I)$ maps a French word f_i in $\mathbf{f}$ to an English word e_{a_i} in $\mathbf{e}$. We have $a_i \in [0, J]$, where $a_i = 0$ is the *null alignment*.

The goal is to learn the probabilistic model of a French sentence given an English sentence $p_\Theta(\mathbf{f}|\mathbf{e})$, where $\Theta = \{p(f|e)\}$ is the set of parameters of the model, f and e being the French and English words appearing in the training dataset.

To learn this model, the **EM algorithm** is used, where the alignment is the latent variable. The **Expectation step** consists in calculating $\mathbb{E}(\text{count}\langle f, e\rangle)$, the expected number of times the French word f is aligned with the English word e. Next, the **Maximization step** updates the parameters of the model using $\hat{p}(f|e) \propto \mathbb{E}(\text{count}\langle f, e\rangle)$.

Once the model has been trained, it can be decoded by calculating the most likely alignment $\hat{\mathbf{a}} = \underset{\mathbf{a}}{\text{argmax}}\ p(\mathbf{a}|\mathbf{f}, \mathbf{e})$ for a pair of translated sentences $(\mathbf{f}, \mathbf{e})$.

IBM1

The IBM1 model assumes that a word of a French sentence has an equal probability of being aligned with any word of the corresponding English sentence (including the null word), i.e.

$$p(a_i = j|i, J, I) = p(a_i = j|J) = \frac{1}{J+1}$$

Therefore, IBM1 exclusively relies on lexical translation; the ordering of the words in sentence is unaccounted for. The decoding of the model is straightforward: the most likely alignment for a pair of translated sentences $(\mathbf{f}, \mathbf{e})$ is given by $\hat{a}_i = \underset{j}{\text{argmax}}\ p(f_i|e_j)$

IBM2

The IBM2 model assumes that alignments of words that are around the same relative place in their respective sentences are more likely, i.e.

$$p(a_i = j|i, J, I) \propto b(|\frac{i}{I} - \frac{j}{J}|) \qquad j \neq 0$$

where $b : \mathbb{R}^+ \to [0, 1]$ is a decreasing function. The probability of the null alignment ($j = 0$) is a fixed parameter p_0. We chose $b : t \mapsto \exp(-\lambda t)$, where λ is a tuning parameter, as in [**?**].

The decoding of the model is still quite simple: the most likely alignment for a pair of translated sentences $(\mathbf{f}, \mathbf{e})$ is given by $\hat{a}_i = \underset{j}{\text{argmax}}\ p(a_i = j|i, J, I)p(f_i|e_j)$.

The assumption in IBM2 is simple enough that the model is very fast to train, but it is still simplifying since it only takes into account the absolute position of words in a sentence. This is why HMM-based alignment models were introduced to account for the relative position of words.

HMM-based alignment models

The HMM-based alignment models were introduced by *Vogel and al.* and aim at modeling the joint probability of a French sentence $\mathbf{f}$, of length J and an alignment $\mathbf{a} \in \{1, \ldots, I\}^J$, given a English sentence $\mathbf{e}$ of length I. Without, loss of generality, this probability can be expressed as (chain-rule and independance of $\mathbf{f}$ and $\mathbf{a}$):

$$\begin{aligned} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) &= \Pr(f_0, a_0|\mathbf{e}) \prod_{j=1}^{J} \Pr(f_j, a_j|f_1^{j-1}, a_1^{j-1}, \mathbf{e}) \\ &= \Pr(f_0|\mathbf{e}) \cdot \Pr(a_0|\mathbf{e}) \\ &\quad \cdot \prod_{j=1}^{J} \Pr(f_j|f_1^{j-1}, a_1^{j-1}, \mathbf{e}) \Pr(a_j|f_1^{j-1}, a_1^{j-1}, \mathbf{e}) \end{aligned}$$

The HMM model reduced the generality of the previous formula by assuming first-order dependance of alignments and translation probability of French word at position j to be only dependent on the English word at position a_j, thus yielding the HMM model:

$$\begin{aligned} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) &= p(f_0|e_{a_0}) p(a_0) \\ &\quad \cdot \prod_{j=1}^{J} p(f_j|e_{a_j}) \cdot p(a_j|a_{j-1}, I) \end{aligned}$$

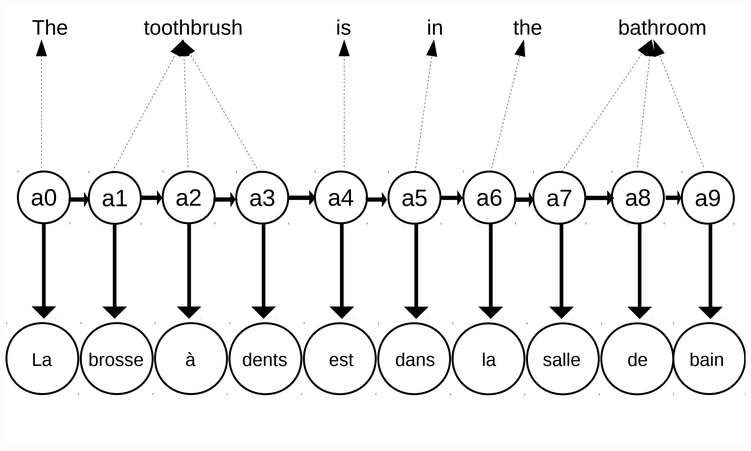

Figure 1: The HMM-based alignment model

Figure **??** presents the graphical model associated with HMM-based alignment models. What can be noticed is that, in this context, each sentence pair represents a Hidden Markov Model which parameters are shared with all other sentences of the corpus. The **parameters for the HMM model** are

$\Theta = \{$

$p(i)$	the initial alignment probabilities ($p(a_0)$)
$p(i\|i', I)$	the transition matrix
$p(f\|e)$	the emission (translation) probabilities$\}$

A further assumption from *Vogel and al.* is that alignment transition probabilities **only depend on relative positions**: $p(i|i', I) = p(i' - i|I)$

Just as for the IBM models, the observations are the sequences of French words in French sentences, while the alignments (and thus the corresponding English words) act as hidden variables. A way to solve this problem is to use an **EM principle**. The **Expectation step** consists in computing alignment unary and binary probabilities in each sentence pair $(\mathbf{f}, \mathbf{e})$.

$$\begin{aligned} \gamma_i(j) &= p(a_j = i|\mathbf{f}) \\ \xi_{k,l}(j) &= p(a_j = k, a_{j+1} = l|\mathbf{f}) \end{aligned}$$

This can be performed using the Forward-Backward algorithm. The **Maximisation step** consists in re-estimating the model's parameters according to the previously computed probabilities.

The **Viterbi** algorithm is then used to retrieve the most likely alignments in the corpus, according to the computed transition and emission probabilities.

Results and discussions

Your text with scientific results or something... Your text with scientific results or something... Your text with scientific results or something...

Your text with scientific results or something... Your text with scientific results or something... Your text with scientific results or something... Your text with scientific results or something...

Example results

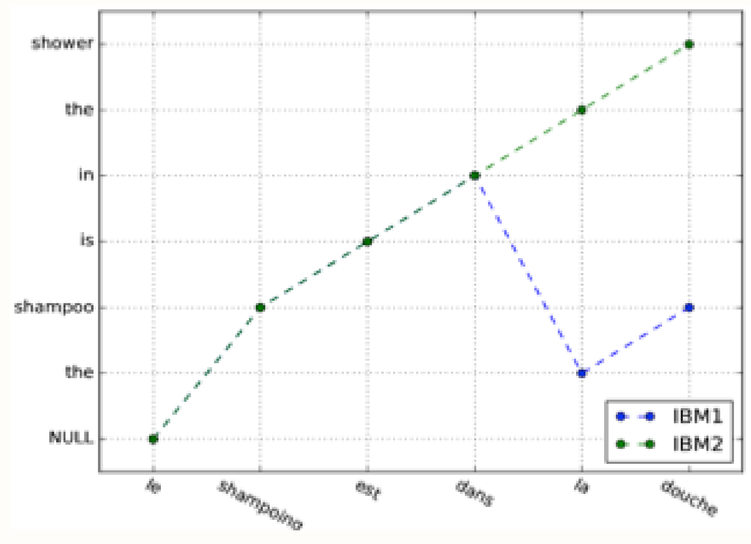

Figure 2: Comparison of the IBM models 1 and 2 example

The above figure shows an example of the decoding results with IBM models 1 and 2 for a same sentence pair. This example illustrates the improvement of IBM2 over IBM1: the English word "the" appears twice in the English sentence, so the IBM1 does not make any distinction between them. However, IBM2 takes the position of words into account, so since the French word "la" is towards the end of the sentence, it is aligned with the second "the".

Summary and conclusions

References

[1] S. Vogel, H. Ney, and C. Tillmann, HMM-based word alignment in statistical translation, 1996

[2] P. Brown et al., The Mathematics of Machine Translation: Parameter Estimation, 1993

[3] C. Dyer et al., A Simple, Fast, and Effective Reparameterization of IBM Model 2, 2013