# A 2D CRF Model for Sentence Alignment

**Yong Xu, François Yvon**
LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay
{yong, yvon}@limsi.fr

### Abstract

The identification of parallel segments in parallel or comparable corpora can be performed at various levels. Alignments at the sentence level are useful for many downstream tasks, and also simplify the identification of finer grain correspondences. Most state-of-the-art sentence aligners are unsupervised, and attempt to infer endogenous alignment clues based on the analysis of the sole bitext. The computation of alignments typically relies on multiple simplifying assumptions, so that efficient dynamic programming techniques can be used. Because of these assumptions, high-precision sentence alignment remains difficult for certain types of corpora, in particular for literary texts. In this paper, we propose to learn a supervised alignment model, which represents the alignment matrix as two-dimensional Conditional Random Fields (2D CRF), converting sentence alignment into a structured prediction problem. This formalism enables us to take advantage of a rich set of overlapping features. Furthermore, it also allows us to relax some assumptions in decoding.

**Keywords:** Sentence Alignment, Conditional Random Fields

## 1. Introduction

The extraction of parallel segments in parallel or comparable corpora can be performed at various levels of granularity (documents, paragraphs, sentence, phrases, chunks, words, etc). For parallel texts or *bitexts*, i.e. pairs of texts assumed to be mutual translations, sentence alignment is a well-defined task in the processing pipeline (Wu, 2010; Tiedemann, 2011). For comparable corpora, sentence alignment techniques are used to mine parallel segments (Munteanu and Marcu, 2005; Uszkoreit et al., 2010). Sentence alignment is used in many applications, such as Statistical Machine Translation (SMT) (Brown et al., 1991), Computer-Assisted Tools, Translator Training (Simard et al., 1993a) and Language Learning (Nerbonne, 2000; Kraif and Tutin, 2011). In SMT, sentence alignment mostly aims at extracting parallel sentence pairs from large-scale corpora (e.g. bilingual parliament proceedings, web-crawled multilingual materials) to fuel downstream statistical processing. For such use, the alignment problem is considered to be solved: on the one hand, it is possible to discard unreliable alignments or difficult pairs (although, as pointed out by Uszkoreit et al. (2010), this might lead to a waste of training material); on the other hand, Goutte et al. (2012) showed that the translation quality of SMT (as measured by BLEU and METEOR) is robust to noise levels of $\approx 30\%$ in sentence alignments.

For other applications, the situation is quite different: First, a requirement may be to align the full bitext, for instance in translation checking (Macklovitch, 1994) or bilingual reading (Pillias and Cubaud, 2015; Yvon et al., 2016). Second, certain types of corpora exhibit important translational irregularities, making high precision alignment difficult. In particular, Yu et al. (2012; Lamraoui and Langlais (2013) showed the link-level F-score of state-of-the-art sentence aligners on bilingual fictions remains unsatisfactory. It was for instance found that the best link-level F-score obtained for “De la Terre à La Lune” (J. Verne), a subpart of the BAF corpus (Simard, 1998), was only around $78\%$.

In this paper, we consider the full sentence alignment problem for difficult bitexts, e.g. literary works and study how supervised learning techniques can help improve this state of affair. More precisely, inspired by the approach of (Mújdricza-Maydt et al., 2013), we propose to represent the alignment matrix by a two-dimensional CRF model, supervised by both reference alignments and external parallel corpora. We use a binary variable to represent the existence of alignment relation between each source and target sentence pair. Once all variables are predicted, we can recover conventional alignment links from the posterior matrix. This representation is very general and dispenses with problematic assumptions, at the cost of a more complex inference procedure.

The rest of this paper is structured as follows. In Section 2., we review some state-of-the-art methods, analyze their limitations and motivate our model. We detail the training and inference in Section 3. Experiments are reported in Section 4. Finally, we conclude and give perspectives for future work in Section 5.

## 2. Motivations

The development of bitext sentence alignment techniques dates back to the early 90s (Brown et al., 1991; Gale and Church, 1991; Simard et al., 1993b; Chen, 1993). Thanks to a sustained research effort, many high-quality aligners are nowadays publicly available, see e.g. (Moore, 2002; Varga et al., 2005; Braune and Fraser, 2010; Lamraoui and Langlais, 2013). A recent evaluation of these tools is in (Xu et al., 2015).

Most state-of-the-art aligners share a two-step approach.[1] A first, relatively coarse decoding pass extracts a set of parallel sentence pairs that the system deems reliable (for instance using length-based information). These pairs serve as either anchor points to reduce the search space of subsequent steps, or as seeds to obtain better parallelism estimation tools (for instance a classifier or a bilingual lexicon), or both. A second decoding pass, using the information gathered during the first step, realigns the bitext. Most of these alignments tools are unsupervised, so that the system has

[1] (Melamed, 1999) is a notable exception.

to collect information from the sole bitext(s) that need to be aligned. In decoding, aligners often make the following assumptions: (a) alignment links lie around the bitext diagonal; (b) there exist limited number of link types. These two assumptions, together with the convention that alignment links are monotone and associate continuous spans,[2] warrant the use of dynamic programming (DP) techniques to perform the search. The resulting alignment tools are often light-weight and efficient, a major requirement if one wishes to process very large bitexts.

Despite their efficiency and good empirical performance on many corpora, existing sentence alignment tools suffer from a number of problems:

- probabilistic alignment models typically assume a fixed prior distribution over link types, as well as specific choices for length distributions (e.g. Gaussian or Poisson). However, Wu (1994) demonstrated that these assumptions could be inaccurate, especially for language pairs that are not closely related;

- as shown in (Yu et al., 2012), DP-based methods often give poor results for *null links*, i.e. links for which one side is empty. Among the five methods compared in this study, only (Melamed, 1999) predicted a similar number of null links as the reference, while others tended to miss a significant portion of them. A possible reason for this problem is the lack of a coherent scoring mechanism which would allow to fairly compare null and non-null links; this especially applies to methods using lexical clues;

- probabilistic alignment models rely on *local* features, and ignore contextual evidences. It might be beneficial to explore structural dependencies in the training;

- the limitation on link types is also overly restrictive. Six main link types are used in most studies: 0:1, 1:0, 1 : 1, 2 : 1, 1 : 2, 2 : 2, and it is a fact that these types rassemble a large majority of links for most text genres. Xu et al. (2015) however report that, in a reference corpus composed of partial sentence alignments for seven literary bitexts, the other types account for approximately 5% of the total number of links, a non-negligible portion for full-text alignment tasks. Besides, such intrinsic model errors can propagate during the DP process.

Inspired by the model of Müjdricza-Maydt et al. (2013), we propose a two-dimensional CRF model for sentence alignment. We use a binary variable to model the existence of the parallelism relation between one source-to-target sentence pair, and include contextual information in our predictions. Decoding consists of classifying each variable as negative or positive. Furthermore, the model structure is richer than that of Müjdricza-Maydt et al. (2013) and includes an explicit representation of null links.

# 3. The 2D CRF Model

## 3.1. The model

Given a sequence of source language sentences $E_1^I = E_1, ..., E_I$ and a sequence of target sentences $F_1^J = F_1, ..., F_J$,[3] we propose a 2D CRF model to predict the presence of link between any pair of sentences $[E_i; F_j]$, where $1 \leq i \leq I, 1 \leq j \leq J$. Note that similar models have also been developed for sub-sentential alignments (Niehues and Vogel, 2008; Cromières and Kurohashi, 2009; Burkett and Klein, 2012). Each pair $[E_i; F_j]$ gives rise to a binary variable $\mathbf{y}_{i,j}$, whose value is 1 (*positive*) if $E_i$ is aligned to $F_j$, and 0 (*negative*) otherwise. For the sequence pair $E_1^I$ and $F_1^J$, there are $I \times J$ such variables, collectively denoted as $\mathbf{y}$. Dependencies between links are modeled as follows. For each pair $[E_i; F_j]$, we assume that the associated variable $\mathbf{y}_{i,j}$ depends on $\mathbf{y}_{i-1,j}$, $\mathbf{y}_{i+1,j}$, $\mathbf{y}_{i,j-1}$, $\mathbf{y}_{i,j+1}$, $\mathbf{y}_{i-1,j-1}$ and $\mathbf{y}_{i+1,j+1}$. In other words, it depends on the presence of links $[E_{i-1}; F_j]$, $[E_{i+1}; F_j]$, $[E_i; F_{j-1}]$, $[E_i; F_{j+1}]$, $[E_{i-1}; F_{j-1}]$ and $[E_{i+1}; F_{j+1}]$. Figure 1 displays a graphical representation of the model.

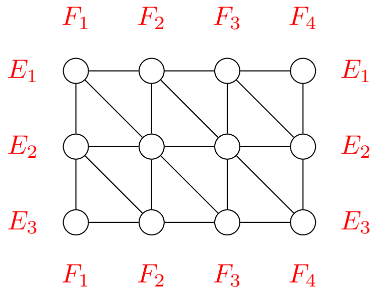


Figure 1: The 2D CRF model, for a bitext of 3 source $E_1 - E_3$ and 4 target sentences $F_1 - F_4$.

The topology of our model differs from the proposal of Müjdricza-Maydt et al. (2013), where each diagonal of the alignment matrix was modeled as a linear chain CRF. This topology captured the important diagonal direction dependency, but did not encode the horizontal or vertical dependencies. Another important difference lies on the generation of final outputs. Müjdricza-Maydt et al. (2013) variable labels to encode the corresponding link type (e.g. 1:1, 2:1). Note that this encoding makes it impossible to include all link types, and has also a bearing on the computational cost, since the inference complexity of a linear chain CRF is quadratic in the number of labels. As a result, these authors only considered 6 link types (1:1, 1:2, 2:1, 1:3, 3:1, $F$).[4] In our model, all prediction variables are binary. We generate final links using the transitive closure operation according to sentence alignment conventions, which can theoretically lead to any possible link type. For instance, an all zero-valued $j^{th}$ column indicates an unaligned target sentence $F_j$; if $\mathbf{y}_{p,q}$ is the only positive value in the $p^{th}$ row and $q^{th}$

[2] If a group of sentences on one side has no correspondence on the other side, they form a null link.

[3] Note we are referring languages as *source* and *target* only for convenience. Thus, *source* does not necessary indicate the original language of the bitext, nor does *target* indicate the translation language.

[4] F stands for all other link types or unaligned sentences.

column, then there is a 1:1 link $[E_p; F_q]$, etc. In fact, the model can express finer correspondences than conventional alignment link representations. For example, if both $E_{u-1}$ and $E_u$ are aligned to $F_{v-1}$, $E_u$ is further aligned to $F_v$, our formalism can represent exactly the relation, while the alignment link representation would contain a coarser 2:2 link $[E_{u-1}, E_u; F_{v-1}, F_v]$.

We use two kinds of clique potentials in our model: *node potentials* and *edge potentials*. We impose that all single node cliques use the same clique template, i.e. they share the same set of feature functions and corresponding weights. For edge potentials, we use distinct clique templates for vertical, horizontal and diagonal edges. One main limitation of this model is that it does not include long distance dependencies, which makes it difficult to encode certain types of constraints (e.g. that alignment links should not cross). The model for a pair of sentence sequences $[E; F]$ (as a shorthand for $[E_1^I; F_1^J]$) can be written as:

$$p(\mathbf{y}|E, F) = \frac{1}{Z(E,F)} \prod_{\nu} \Phi_n(\mathbf{y}_\nu)\Phi_v(\mathbf{y}_\nu)\Phi_h(\mathbf{y}_\nu)\Phi_d(\mathbf{y}_\nu)$$

where $\nu \in \{(i,j) : 1 \leq i \leq I, 1 \leq j \leq J\}$, $\Phi_n(\mathbf{y}_\nu)$ stands for the single node potential at $\nu$, $\Phi_v(\mathbf{y}_\nu)$ represents the potential on the vertical edge connecting $\nu$ and the node just below it:

$$\forall j, \Phi_v(\mathbf{y}_{i,j}) = \begin{cases} 1 & \text{if} \quad i = I \\ \Phi_v(\mathbf{y}_{i,j}, \mathbf{y}_{i+1,j}) & \text{if} \quad 1 \leq i < I \end{cases}$$

$\Phi_h(\mathbf{y}_\nu)$ (the horizontal potential) and $\Phi_d(\mathbf{y}_\nu)$ (the diagonal potential) are defined similarly. $Z(E,F) = \sum_{\mathbf{y}'} \prod_\nu \Phi_n(\mathbf{y}'_\nu)\Phi_v(\mathbf{y}'_\nu)\Phi_h(\mathbf{y}'_\nu)\Phi_d(\mathbf{y}'_\nu)$ is the normalization factor (*the partition function*) of the CRF. All potentials take the generic form of a log-linear combination of feature functions:

$$\Phi_\nu(\mathbf{y}_\nu) = \exp\{\boldsymbol{\theta}^\top \mathbf{F}_\nu(\mathbf{y}_\nu)\},$$

where $\mathbf{F}_\nu$ and $\boldsymbol{\theta}$ are the feature and weight vectors. We also use $\ell^2$ regularization with scaling parameter $\alpha > 0$.[5]

### 3.2. Learning the 2D CRF model

The conventional learning criteria for CRF is the Maximum Likelihood Estimation (MLE). For a set of fully observed training instances $\mathcal{A} = \{(E^{(s)}, F^{(s)}, \mathbf{y}^{(s)})\}$, MLE consists of maximizing the log-likelihood of the training set with respect to model parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \alpha\}$. The log-likelihood is concave with respect to the weight vector, which warrants the use of convex optimization techniques to obtain parameter estimates. In order to do this, we need the gradient of the likelihood function with respect to the weight vector.

Computing the gradients requires two kinds of marginal probabilities: single node marginals $p(\mathbf{y}_{i,j}|E^{(s)}, F^{(s)})$ and edge marginals $p(\mathbf{y}_{i,j}, \mathbf{y}_{i+1,j}|E^{(s)}, F^{(s)})$, $p(\mathbf{y}_{i,j}, \mathbf{y}_{i,j+1}|E^{(s)}, F^{(s)})$, and $p(\mathbf{y}_{i,j}, \mathbf{y}_{i+1,j+1}|E^{(s)}, F^{(s)})$. We need to perform inference to compute these marginals. Since the topology of our model contains loops, we use the *Loopy Belief Propagation* (LBP) inference algorithm. Even though LBP is an *approximate* inference algorithm with no convergence guarantee, Murphy et al. (1999) observe that it often gives reasonable estimates (assuming it converges).

For a tree-structured undirected graphical model, the message from a node $\mathbf{y}_\mu$ to a neighboring node $\mathbf{y}_\nu$ takes the following form (Wainwright and Jordan, 2008):

$$m_{\mu\nu}(\mathbf{y}_\nu) \propto \sum_{\mathbf{y}_\mu} \Phi(\mathbf{y}_\mu)\Phi(\mathbf{y}_\nu, \mathbf{y}_\mu) \prod_{\gamma \in N(\mu)\backslash\nu} m_{\gamma\mu}(\mathbf{y}_\mu)$$

where $N(\mu)$ denotes the set of neighbors of $\mu$. LBP is performing such message passing procedure on a cyclic graph. Once message passing has converged, the single node and edge marginals (a.k.a. "beliefs") are expressed as:

$$b_\nu(\mathbf{y}_\nu) \propto \Phi(\mathbf{y}_\nu) \prod_{\gamma \in N(\nu)} m_{\gamma\nu}(\mathbf{y}_\nu)$$

$$b_{\mu\nu}(\mathbf{y}_\mu, \mathbf{y}_\nu) \propto \Phi(\mathbf{y}_\mu)\Phi(\mathbf{y}_\nu)\Phi(\mathbf{y}_\nu, \mathbf{y}_\mu) \prod_{\delta \in N(\mu)\backslash\nu} m_{\delta\mu}(\mathbf{y}_\mu) \prod_{\gamma \in N(\nu)\backslash\mu} m_{\gamma\nu}(\mathbf{y}_\nu)$$

In practice, it is possible that LBP does not converge for certain training instances. In this case, we simply stop it after 100 iterations. Convex optimization routines also require to compute the log-partition function $\log Z(E,F)$, as a part of the likelihood function. LBP approximates this quantity with the Bethe Free Energy (Yedidia et al., 2001).

In learning, we first train the CRF without any edge potential (thus making the model similar to the simpler MaxEnt model), and use it to initialize the parameter vector of node potentials. We then randomly initialize other parameters,[6] and use the L-BFGS algorithm (Liu and Nocedal, 1989) implemented in the SciPy package to perform parameter learning, this time with all potentials.

### 3.3. Search in the 2D CRF model

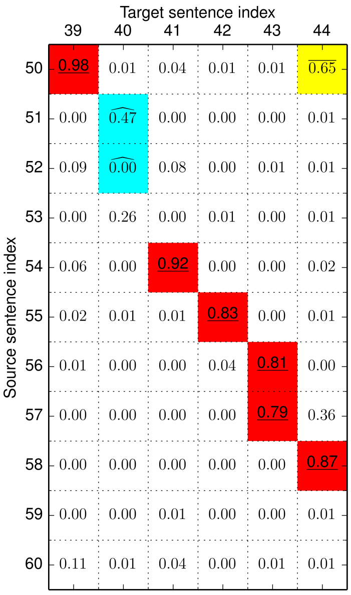
For the 2D CRF model, we perform the search in multiple steps. First, we run the BMA algorithm (Moore, 2002) to extract high-confidence 1:1 links. This algorithm first extracts reliable 1:1 sentence pairs from a bitext, using only length information, then trains a small IBM Model 1 based on these links, finally realigns the bitext using both length and lexical information. It returns a set of $1:1$ sentence pairs. As reported in (Yu et al., 2012), BMA tends to obtain a very good precision, at the expense of a less satisfactory recall. Furthermore, BMA computes posterior probabilities for every possible link, which are then used as confidence scores. We filter the result links with a very high posterior probability threshold ($\geq 0.99999$) (this threshold is much higher than BMA's default choice). These links segment the entire search space into sub-blocks. For each sub-block, we construct a 2D CRF model, and perform decoding. As exact Maximum A Posteriori decoding is intractable, instead, we run max-product LBP independently, and pick the *local best* label for each node. The label assigned to a variable $\mathbf{y}_{i,j}$ is

$$\arg\max_{l \in \{0,1\}} b_{i,j}(l)$$

[5] In the experiments, $\alpha$ is tuned on a development set, and takes the value 0.1.

[6] See (Sutton, 2008, 88–89) for a discussion on parameter initialization of general CRFs trained using LBP.

This procedure returns a set of *sentence-level* links. Since the sizes of the sub-blocks are often small (generally smaller than $10 \times 10$), decoding is very fast in practice.
Figure 2 displays an *alignment prediction matrix*.[7] It contains four types of cells, corresponding to four types of predictions: true positive (red, with underlined score), true negative (white, with normal score), false positive (yellow, with overlined score), false negative (cyan, with hatted score). The score in each cell is the marginal probability of the pair being positive, as computed by the CRF. A red or yellow cell indicates a sentence-level link predicted by the model.



|  | 39 | 40 | 41 | 42 | 43 | 44 |
|---|---|---|---|---|---|---|
| 50 | $\underline{0.98}$ | 0.01 | 0.04 | 0.01 | 0.01 | $\overline{0.65}$ |
| 51 | 0.00 | $\widehat{0.47}$ | 0.00 | 0.00 | 0.00 | 0.01 |
| 52 | 0.09 | $\widehat{0.00}$ | 0.08 | 0.00 | 0.01 | 0.01 |
| 53 | 0.00 | 0.26 | 0.00 | 0.01 | 0.00 | 0.01 |
| 54 | 0.06 | 0.00 | $\underline{0.92}$ | 0.00 | 0.00 | 0.02 |
| 55 | 0.02 | 0.01 | 0.01 | $\underline{0.83}$ | 0.00 | 0.01 |
| 56 | 0.01 | 0.00 | 0.00 | 0.04 | $\underline{0.81}$ | 0.00 |
| 57 | 0.00 | 0.00 | 0.00 | 0.00 | $\underline{0.79}$ | 0.36 |
| 58 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | $\underline{0.87}$ |
| 59 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 |
| 60 | 0.11 | 0.01 | 0.04 | 0.00 | 0.01 | 0.01 |

Figure 2: An alignment prediction matrix.

Two types of errors exist in the alignment prediction matrix: false negatives (cyan cells with hatted scores) and false positives (yellow cells with overlined scores). We cannot easily deal with false negatives. False positives introduce noises, for example, the pair $(50, 44)$ in Figure 2 (the upper right corner). The two positive pairs $(50, 39)$ and $(50, 44)$ lead to two separate links involving the same source sentence, which violates the general convention of sentence alignment. In fact, the pair $(50, 44)$ is clearly wrong: it links the first source sentence with the last target one, thus overlapping with all other positive sentence-level links.[8] In our experiments, in all sub-blocks, true positive sentence-level links always lie around the main diagonals. We have used the following heuristics to smooth the alignment prediction matrix:

1. perform a linear regression on all predicted positive sentence-level links, then take a band of fixed width around the regression line, and drop positive links that lie outside of this band.[9]

2. if after this step, there are still separate links involving the same sentence, we take the positive sentence-level links in the surrounding window with width 5, and discard the ones which are inconsistent with the surrounding links;

3. if it is still undecidable, we perform again a linear regression of positive sentence-level links in the surrounding window, and discard the link that is farthest away from the regression line.

In practice, step 3 was hardly performed.
Finally, to turn sentence-level links into *alignment-level* links, we apply the following rules:

1. consecutive sentence-level links in the horizontal or vertical directions are combined into a large alignment-level link;

2. a sentence-level link without horizontal or vertical neighbors becomes a 1:1 type alignment-level link.

These rules follow from the interpretation of our model, where an $n{:}m$ type alignment-level link decomposes into $n * m$ sentence-level links.

# 4. Experiments

## 4.1. Features

In the 2D CRF model, feature functions take the form $f(E_1^I, F_1^J, i_1, j_1, i_2, j_2, \mathbf{y}_{i_1,j_1}, \mathbf{y}_{i_2,j_2})$, where $E_1^I$ is the source sequence of $I$ sentences, $F_1^J$ the target sequence of $J$ sentences, $(i_1, j_1)$ and $(i_2, j_2)$ neighboring source-target indices, $\mathbf{y}_{i_1,j_1}$ and $\mathbf{y}_{i_2,j_2}$ respectively corresponding labels (0 or 1). For each pair $(E_i, F_j)$, we compute the following set of features:

1. The *length difference ratios*. We first compute

$$r_1 = \frac{|len(E_i) - len(F_j)|}{len(E_i)}, \; r_2 = \frac{|len(E_i) - len(F_j)|}{len(F_j)}$$

where the $len()$ function returns the number of characters in one string. Both $r_1$ and $r_2$ are rounded into the interval $[0, 1]$, then discretized into 10 indicator features. This family thus contains 20 features.

2. The *ratio of identical tokens*. Let the function $token()$ return the number of tokens in a string. We count the number of shared tokens in $E_i$ and $F_j$, denote the count by $s$, compute two ratios $\frac{s}{token(E_i)}$ and $\frac{s}{token(F_j)}$, then discretize each into 10 features.

3. The *relative index difference*. We discretize the quantity $|\frac{i}{I} - \frac{j}{J}|$ into 10 features.

[7] Note these matrices are drawn just after the CRF decoding, before the post-processing described below.

[8] Note that this particular matrix was computed by an early version of the 2D CRF model. We show it here for illustration purpose. In later versions, the model is augmented with features capturing the relative position information, which effectively prevents this kind of errors.

[9] The band width is taken to be half of the number of sentences of the shorter one of the two sides.

| Book | # Links | # Sent_EN | # Sent_FR |
|---|---|---|---|
| Alice's Adventures in Wonderland | 746 | 836 | 941 |
| Candide | 1,230 | 1,524 | 1,346 |
| Vingt Mille Lieues sous les Mers | 778 | 820 | 781 |
| Voyage au Centre de la Terre | 714 | 821 | 754 |
| *Total* | 3,468 | 4,001 | 3,822 |

Table 1: The training corpus of the 2D CRF model.

| Book | # Links | # Sent_EN | # Sent_FR |
|---|---|---|---|
| De la Terre à la Lune (BAF) | 2,520 | 2,554 | 3,319 |
| Du Côté de chez Swann | 463 | 495 | 492 |
| Emma | 164 | 216 | 160 |
| Jane Eyre | 174 | 205 | 229 |
| La Faute de l'Abbe Mouret | 222 | 226 | 258 |
| Les Confessions | 213 | 236 | 326 |
| Les Travailleurs de la Mer | 359 | 389 | 405 |
| The Last of the Mohicans | 197 | 205 | 232 |
| *Total of* Manual en-fr | 1,792 | 1,972 | 2,102 |

Table 2: The test corpus, made of the literary part of BAF and the manual en-fr corpus.

4. The *lexical translation scores*. Let $token(E_i) = m$ and $token(F_j) = n$, we compute the IBM Model 1 scores:

$$\begin{aligned} T_1(E_i, F_j) &= \frac{1}{n}\sum_{s=1}^{n} \log\left(\frac{1}{m} * \sum_{k=1}^{m} p(F_{js}|E_{ik})\right) \\ T_2(E_i, F_j) &= \frac{1}{m}\sum_{k=1}^{m} \log\left(\frac{1}{n} * \sum_{s=1}^{n} p(E_{ik}|F_{js})\right) \end{aligned}$$

   where $F_{js}$ is the $s^{th}$ token of $F_j$. The lexical translation probabilities $p$ are computed using an IBM 1 model trained on the EN-FR Europarl corpus (Koehn, 2005). After discretizing $T_1$ and $T_2$, we obtain 10 features for each alignment direction.

5. The *span coverage*. We split a string into several **spans** by segmenting on punctuations (except for the quotation marks). For each source span $span\_e$, we compute the translation score $T_2(span\_e, F_j)$. If the score is larger than a threshold,[10] we consider $span\_e$ as being **covered**. We then compute the ratio of covered source spans and the ratio of covered target spans, and discretize each into 10 features.

6. The *label transition*. These features capture the regularity of the transition of labels from one node $(E_i, F_j)$ to one of its neighbors (e.g. $(E_{i+1}, F_j)$). For each of the three types of neighbors (vertical, horizontal, diagonal), we define four label transition features (because our prediction variables are binary). For example, for the vertical template, we define

$$\begin{aligned} g_{00}(i,j) &= \delta\{\mathbf{y}_{i,j} = 0 \wedge \mathbf{y}_{i+1,j} = 0\} \\ g_{01}(i,j) &= \delta\{\mathbf{y}_{i,j} = 0 \wedge \mathbf{y}_{i+1,j} = 1\} \\ g_{10}(i,j) &= \delta\{\mathbf{y}_{i,j} = 1 \wedge \mathbf{y}_{i+1,j} = 0\} \\ g_{11}(i,j) &= \delta\{\mathbf{y}_{i,j} = 1 \wedge \mathbf{y}_{i+1,j} = 1\} \end{aligned}$$

   where $\delta$ is the Kronecker delta function. We have similar features for horizontal and diagonal transitions. In total, this family contains 12 features.

7. The *augmented length difference ratio*. This family only applies to the vertical and horizontal edge potentials, under the condition that the two neighboring pairs are both positive. In the vertical (resp. horizontal) case, we combine the two consecutive source (resp. target) sentences $E_i, E_{i+1}$ (resp. $F_j, F_{j+1}$) into one new sentence $E'$ (resp. $F'$), then apply the computations carried out for feature family 1 for the pair $(E', F_j)$ (resp. $(E_i, F')$).

8. The *augmented translation score*. This family only applies to vertical and horizontal edge potentials, under the condition that the two neighboring pairs are both positive. We construct $E'$ (resp. $F'$) as in the previous feature family. We then compute the augmented translation score $T_1(E', F_j) - T_1(E_i, F_j)$ (resp. $T_2(E_i, F') - T_2(E_i, F_j)$). The intuition is that a longer partial translation is better than a shorter one. Each score is discretized into 10 features.

Note feature families 6, 7 and 8 are computed only when possible. Feature families 5, 7 and 8 are new in our model. Others have been used in previous methods, for instance, (Munteanu and Marcu, 2005; Yu et al., 2012; Tillmann and Hewavitharana, 2013; Mújdricza-Maydt et al., 2013).

## 4.2. Learning corpus

The training of the 2D CRF model requires reference alignments. We have used the reference sentence alignments collected for an ongoing project.[11] The training corpus contains alignment links of four books: "Alice's Adventures in Wonderland" (L. Carroll), "Candide" (Voltaire), "Vingt

[10] In our experiments, the threshold is set to $\log(1e-3)$

[11] See http://transread.limsi.fr, where most textual resources can be downloaded.

| | Sentence level F-score | | | | | | |
|---|---|---|---|---|---|---|---|
| | GMA | BMA | Hunalign | Garg | Yasa | MaxEnt | CRF |
| De la Terre à la Lune (BAF) | 72.9 | 77.3 | 81.9 | 77.3 | 86.2 | 76.6 | 84.0 |
| Du Côté de chez Swann | 95.4 | 88.9 | 89.4 | 95.0 | 95.2 | 96.0 | 94.3 |
| Emma | 73.8 | 52.1 | 62.8 | 61.2 | 73.8 | 71.2 | 69.4 |
| Jane Eyre | 88.0 | 54.6 | 59.4 | 84.2 | 82.5 | 88.0 | 77.2 |
| La Faute de l'Abbé Mouret | 94.8 | 83.8 | 82.8 | 98.7 | 97.7 | 98.9 | 90.8 |
| Les Confessions | 82.8 | 49.9 | 48.5 | 80.5 | 82.8 | 86.1 | 76.6 |
| Les Travailleurs de la Mer | 87.8 | 79.6 | 78.8 | 91.5 | 90.4 | 91.9 | 89.1 |
| The Last of the Mohicans | 94.9 | 76.0 | 77.0 | 95.6 | 94.5 | 95.0 | 91.1 |
| *Average on* `manual en-fr` | 88.2 | 69.3 | 71.2 | 86.7 | 88.1 | 89.6 | 84.1 |

Table 3: Sentence level F-scores of the 2D CRF method on the test corpus, compared with state-of-the-art methods.

| | BMA | | | MaxEnt | | | CRF | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F |
| De la Terre à la Lune (BAF) | 97.2 | 64.1 | 77.3 | 72.0 | 81.8 | 76.6 | 95.5 | 74.9 | 84.0 |
| Du Côté de chez Swann | 99.5 | 80.3 | 88.9 | 97.1 | 94.9 | 96.0 | 96.3 | 92.5 | 94.3 |
| Emma | 89.8 | 36.7 | 52.1 | 62.8 | 82.3 | 71.2 | 76.1 | 63.7 | 69.4 |
| Jane Eyre | 93.7 | 38.5 | 54.6 | 86.7 | 89.3 | 88.0 | 86.6 | 69.6 | 77.2 |
| La Faute de l'Abbé Mouret | 99.5 | 72.3 | 83.8 | 98.9 | 98.9 | 98.9 | 98.2 | 84.5 | 90.8 |
| Les Confessions | 98.4 | 33.4 | 49.9 | 89.3 | 83.2 | 86.1 | 92.6 | 65.4 | 76.6 |
| Les Travailleurs de la Mer | 97.7 | 67.2 | 79.6 | 90.8 | 93.0 | 91.9 | 97.1 | 82.2 | 89.1 |
| The Last of the Mohicans | 98.7 | 61.8 | 76.0 | 94.2 | 95.8 | 95.0 | 97.1 | 85.7 | 91.1 |
| *Average on* `manual en-fr` | 96.8 | 55.7 | 69.3 | 88.5 | 91.1 | 89.6 | 92.0 | 77.7 | 84.1 |

Table 4: The comparison of BMA, MaxEnt and the 2D CRF model, using sentence-level measures. P stands for Precision, R is Recall, and F is F-score.

Mille Lieues sous les Mers", and "Voyage au Centre de la Terre" (both by J. Verne). Table 1 displays the statistics of the training corpus.

We have to convert the training corpus into a training set. A training instance is a fully observed sentence-level alignment matrix. In order to make train conditions as close as possible to test conditions, each fully aligned book was segmented into sub-blocks, again using high confidence 1:1 links computed by BMA as anchor points. Each sub-block, annotated with reference alignments, is then turned into one training instance. This strategy has the additional benefit to greatly reduce the total number of prediction variables, hence make the training less memory consuming. Besides, the training can enjoy better parallelization. There is potentially another advantage of using smaller training instances. Since our model contains one predictive variable for each pair of source-target sentences, there are roughly quadratically many negative examples, and linearly many positive ones. This data unbalance problem becomes more severe as the size of the prediction matrix grows larger. Using smaller training instances helps alleviate this problem.

Using this strategy, we obtain 450 fully observed alignment matrices. We use 360 for the training set, 90 as the development set. Among the 7,095 labeled sentence pairs, approximately 77% are negative.

For the test, we use the fully aligned novel "De la Terre à la Lune" in the BAF corpus and the `manual en-fr` corpus composed of 7 partial alignments of literary bitexts. Table 2 gives the statistics of the test corpus. Recall our first step is to use filtered results of BMA as anchor points to segment the search space. With the filtering threshold 0.99999, the anchor point precision is 0.89 on "De la Terre à la Lune", and 0.96 on the `manual en-fr` corpus.

### 4.3. Results

We evaluate alignment results at two levels of granularity: the alignment level and the sentence level. At the alignment level, a link in the output alignment is considered correct if exact the same link is also in the reference alignment. At the sentence level, we decompose a $m{:}n$ type link in the reference alignment into $m \times n$ sentence pairs, all considered as correct. The same decomposition applies to computed links. We summarize precision and recall ratios into F-scores.

Since the 2D CRF model is intrinsically trained to optimize **sentence-level** metrics, we first look at its sentence-level performance, summarized in Table 3. For the sake of comparison, we also display the performance of six other state-of-the-art aligners: GMA (Melamed, 1999), BMA, Hunalign (Varga et al., 2005), Garg (as shorthand for Gargantua) (Braune and Fraser, 2010), Yasa (Lamraoui and Langlais, 2013), MaxEnt (Xu et al., 2015). The CRF model achieves great improvements over BMA and Hunalign. Its average score on the `manual en-fr` corpus is slightly inferior to other systems, but it obtains the second best F-measure on the large bi-text "De la Terre à la Lune". We note that Yasa, perhaps the most lightweight tool, is very robust with respect to the sentence-level measure.

The first decoding step of both MaxEnt and CRF uses a subset of BMA's results as anchors to segment the bi-text space. Table 4 compares in more detail the performance of these three methods. (Yu et al., 2012; Lamraoui and Langlais, 2013) have reported that BMA usually delivers very high precision 1 : 1 links. We observe the 2D CRF model preserves a high sentence level precision, and greatly increases the recall. Thus, the 2D CRF model manages to extract true positive sentence pairs from the gaps defined by BMA's links with a very high accuracy. The behavior of MaxEnt varies on different corpus. On the `Manual en-fr` corpus, while it slightly decreases the precision, it obtains the best recall, leading to the best overall performance. However, on "De la Terre à la Lune", its precision is too low compared to BMA and CRF, thus its F-score is worse.

| | Alignment level F-score | | |
|---|---|---|---|
| | BMA | MaxEnt | CRF |
| De la Terre à la Lune (BAF) | 73.6 | 66.5 | 73.3 |
| Du Côté de chez Swann | 91.5 | 93.3 | 90.9 |
| Emma | 57.4 | 51.0 | 55.4 |
| Jane Eyre | 61.1 | 78.9 | 63.2 |
| La Faute de l'Abbé Mouret | 88.4 | 98.0 | 82.8 |
| Les Confessions | 59.6 | 74.0 | 58.1 |
| Les Travailleurs de la Mer | 83.4 | 85.3 | 83.0 |
| The Last of the Mohicans | 82.7 | 90.1 | 84.3 |
| *Average on* `manual en-fr` | 74.9 | 81.5 | 74.0 |

Table 5: Alignment level F-scores of the 2D CRF model, compared with BMA and MaxEnt.

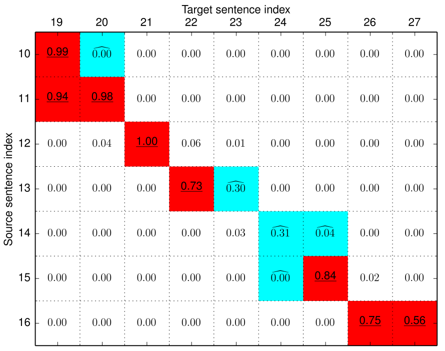
The alignment level F-scores of the CRF model are in Table 5.[12] The CRF achieves comparable alignment level F-scores to BMA on both sub-corpus. Although their average scores on `manual en-fr` are worse than MaxEnt, they outperform it considerably on those more difficult bitexts: "De la Terre à la Lune" and "Emma". In our opinion, this calls for further analyses for the *deployment* of alignment methods: for sentence alignment, it might be beneficial to investigate which types of methods tend to perform well for which types of bitexts, identify indicative characteristics (of methods and bitexts), and deduce operational guidelines.[13] Table 4 and Table 5 together show that, while the 2D CRF model obtains much higher sentence level F-scores than BMA (approximately 15 points on average on `manual en-fr`), their alignment level F-scores are actually comparable. In other words, the CRF does find more true positive sentence pairs, but not all of them contribute to form true links. Take for instance the 2 : 2 link $(14, 15; 24, 25)$ in Figure 3. To correctly recover this link, it is necessary to find at least three among the four cells. Even though the CRF finds one cell $(15; 25)$, this only yields a wrong 1 : 1 link, which, for the alignment level F-score metric, is no better than not finding any pair. While this imbalance between the alignment level and sentence level F-scores can seem surprising, it is by no means uncommon. In fact, this phenomenon was the reason that sentence-level F-score was proposed as an evaluation metric for sentence alignment in (Langlais et al., 1998). Nonetheless, this reinforces our belief that the deployment strategy of alignment methods, as well as evaluation metrics, needs further study.

## 4.4. Analysis

**Error distribution by link type** To better understand the behavior of the 2D CRF model, we perform an error analysis of its results on the `manual en-fr` corpus, with respect to link types. The corresponding statistics are in Table 7. We compare CRF with the MaxEnt approach, which gives the best average score on this corpus.

| Link type | in Ref. | Error MaxEnt | Error CRF |
|---|---|---|---|
| 0:1 | 20 | 18 | 15 |
| 1:0 | 21 | 18 | 15 |
| 1:1 | 1,366 | 105 | 64 |
| 1:2 | 179 | 36 | **98** |
| 1:3 | 32 | 9 | **29** |
| 2:1 | 96 | 32 | **54** |
| 2:2 | 24 | 19 | 20 |
| others | 27 | 15 | 26 |
| *Total* | 1,765 | 252 | 321 |

Table 7: Analyses of the errors of the MaxEnt and the CRF by link type, relative to the number of reference links (in Ref.), for the `manual en-fr` corpus. For example, 20 0 : 1 links are in the reference, and MaxEnt missed 18 of them. Only the link types occurring more than 5 times are reported. This filters out 27 links out of 1,792.



| | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
|---|---|---|---|---|---|---|---|---|---|
| 10 | $\underline{0.99}$ | $\widehat{0.00}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 11 | $\underline{0.94}$ | $\underline{0.98}$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 12 | 0.00 | 0.04 | $\underline{1.00}$ | 0.06 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| 13 | 0.00 | 0.00 | 0.00 | $\underline{0.73}$ | $\widehat{0.30}$ | 0.00 | 0.00 | 0.00 | 0.00 |
| 14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | $\widehat{0.31}$ | $\widehat{0.04}$ | 0.00 | 0.00 |
| 15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | $\widehat{0.00}$ | $\underline{0.84}$ | 0.02 | 0.00 |
| 16 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | $\underline{0.75}$ | $\underline{0.56}$ |

Figure 3: An alignment prediction matrix for a passage of "Les Confessions".

Compared to the MaxEnt method, CRF has a higher recall on null and 1:1 links. Its main weakness lies in the prediction of 1:$n$ and $n$:1 links. After a closer study of the erroneous instances, we find a common pattern of error: when predicting a $m$:$n$ link with $m * n > 1$ (that is, a 1-to-many or many-to-many link), the CRF often correctly labels some sentence pairs as positive, while leaving others as negative. Figure 3 displays an alignment prediction matrix for a pas-

[12] We only show BMA, MaxEnt and CRF in this table, since (Xu et al., 2015) reported MaxEnt obtained the best average alignment F-score on the `manual en-fr` corpus.

[13] This is in line with the views of Deng et al. (2007) and Lamraoui and Langlais (2013), who suggested to model sentence alignment as part of the target application, so that it can benefit the optimization conducted toward the task.

| | 2D CRF | | | MaxEnt | | |
|---|---|---|---|---|---|---|
| | #Null (in Ref.) | #Null (in Hyp.) | #Correct | #Null (in Ref.) | #Null (in Hyp.) | #Correct |
| De la Terre à la Lune (BAF) | 714 | 1,311 | 672 | 714 | 150 | 91 |
| Du Côté de chez Swann | 9 | 27 | 8 | 9 | 5 | 3 |
| Emma | 41 | 85 | 28 | 41 | 2 | 2 |
| Jane Eyre | 10 | 77 | 7 | 10 | 0 | 0 |
| La Faute de l'Abbé Mouret | 2 | 52 | 2 | 2 | 1 | 1 |
| Les Confessions | 11 | 96 | 11 | 11 | 4 | 2 |
| Les Travailleurs de la Mer | 5 | 78 | 3 | 5 | 2 | 0 |
| The Last of the Mohicans | 12 | 37 | 3 | 12 | 2 | 2 |

Table 6: Performance of the 2D CRF model and the MaxEnt model on predicting null sentences. “#Null in Ref.” is the number of unaligned sentences in the reference alignment; “#Null in Hyp.” is the number of unaligned sentences in the hypothesis alignment computed by the model;“‘#Correct” is the number of correctly predicted null sentences.

sage of Jean-Jacques Rousseau's “Les Confessions”. The corresponding text (correctly aligned) is displayed in Table 8 in the appendix. The CRF fails to predict the 1:2 link $(13; 22, 23)$, only labelling $(13; 22)$ as positive; nor does it find the 2:2 link $(14, 15; 24, 25)$.

The failures of the 2D CRF model on 1-to-many and many-to-many links makes it necessary to study edge potentials. One of the reasons of using a CRF model is its ability to encode the dependencies between neighboring links, with which we expect to better predict non 1:1 links. An obvious direction to investigate is to add more edge features. Current edge features (families 6, 7 and 8) are quite general. It might be helpful to add features that encode finer level clues to edge potentials, e.g. word alignment information.

Besides of features of edge potentials, it might also be possible to consider other alignment matrix decoding algorithms. Compared to our approach, MaxEnt has the advantage of directly scoring alignment-level links, rather than doing it obliquely through sentence-level ones. This is also possible in the 2D CRF model, since LBP can readily compute marginals over edges, or even larger factors. We might use such marginals to improve our post-processing routines.

**Null sentences** Another motivation for the 2D CRF model is that it provides a mechanism where null and non-null links are handled coherently. We summarize its performance for null sentences in Table 6, again, comparing it with the MaxEnt method.

Although the 2D CRF model incorrectly labels many sentences as unaligned, it is indeed able to find the majority of true null sentences, except for “The Last of the Mohicans”. This is where our model seems to be improving, especially when compared to MaxEnt.

## 5. Conclusion

In this paper, we reviewed state-of-the-art sentence alignment methods, identified several recurring problems, and have accordingly proposed a two-dimensional Conditional Random Fields model for the full text sentence alignment task. Our model is theoretically attractive, since it avoids several risky assumptions, computes posterior probabilities for all sentence alignment links, thereby explicitly representing null links, and warrants structured learning of parallelism scores.

In the light of our experimental results and analyses, we conclude that there is clear room of improvement for our 2D CRF model. Currently, while the model is effective at identifying true $1:1$ links with better recall than BMA's, its performance as measured by alignment level metric still needs to be improved. As perspectives, we would like to study the following improvements:

- enforce edge features: current edge features do not seem to be strong enough to balance our rich set of node features. Including features informed with simple word alignment information, such as fertilities and linked regions, seems an obvious way to go;
- add node features that encode the decisions of other systems, e.g. BMA;
- explore ways to simulate a DP process using marginals of edges or larger factors, which might help improve our alignment matrix decoding algorithm.

In the long term, we would like to study ways to characterize tasks and alignment methods, such that it is possible to choose adequate alignment algorithms for specific task requirements.

## 6. Acknowledgements

This work was partly supported by French National Research Agency under project TransRead (ANR-12-CORD-0015).

## 7. Bibliographical References

Braune, F. and Fraser, A. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Coling 2010: Posters*, pages 81–89.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of ACL*, pages 169–176.

Burkett, D. and Klein, D. (2012). Fast inference in phrase extraction models with belief propagation. In *Proceedings of NAACL: HLT*, pages 29–38.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of ACL*, pages 9–16.

Cromières, F. and Kurohashi, S. (2009). An alignment algorithm using belief propagation and a structure-based distortion model. In *Proceedings of EACL*, pages 166–174.

Deng, Y., Kumar, S., and Byrne, W. (2007). Segmentation and alignment of parallel text for statistical machine translation. *Natural Language Engineering*, 13(03):235–260.

Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of ACL*, pages 177–184.

Goutte, C., Carpuat, M., and Foster, G. (2012). The impact of sentence alignment errors on phrase-based machine translation performance. In *Proceedings of AMTA*.

Kraif, O. and Tutin, A. (2011). Using a bilingual annotated corpus as a writing aid: An application for academic writing for EFL users. In *Corpora, Language, Teaching, and Resources: From Theory to Practice. Selected papers from TaLC7*.

Lamraoui, F. and Langlais, P. (2013). Yet Another Fast, Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment? In *Proceedings of MT Summit*, pages 77–84.

Langlais, P., Simard, M., and Véronis, J. (1998). Methods and practical issues in evaluating alignment techniques. In *Proceedings of ACL-COLING*, pages 711–717.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528.

Macklovitch, E. (1994). Using bi-textual alignment for translation validation: the TransCheck system. In *Proceedings of AMTA*, pages 157–168.

Melamed, I. D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25:107–130.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of AMTA*, Lecture Notes in Computer Science 2499, pages 135–144.

Mújdricza-Maydt, E., Köerkel-Qu, H., Riezler, S., and Padó, S. (2013). High-precision sentence alignment by bootstrapping from wood standard annotations. *The Prague Bulletin of Mathematical Linguistics*, (99):5–16.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Murphy, K. P., Weiss, Y., and Jordan, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of UAI*, pages 467–475.

Nerbonne, J., (2000). *Parallel Texts in Computer-Assisted Language Learning*, chapter 15, pages 354–369. Text Speech and Language Technology Series.

Niehues, J. and Vogel, S. (2008). Discriminative word alignment via alignment matrix modeling. In *Proceedings of WMT*, pages 18–25.

Pillias, C. and Cubaud, P. (2015). Bilingual reading experiences: What they could be and how to design for them. In *Proceedings of INTERACT 2015*, pages 531–549.

Simard, M., Foster, G., and Perrault, F. (1993a). Transsearch: A bilingual concordance tool. Technical report, Centre for Information Technology Innovation.

Simard, M., Foster, G. F., and Isabelle, P. (1993b). Using cognates to align sentences in bilingual corpora. In *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research*, pages 1071–1082.

Simard, M. (1998). The BAF: a corpus of English-French bitext. In *Proceedings of LREC*, pages 489–494.

Sutton, C. (2008). *Efficient Training Methods for Conditional Random Fields*. Ph.D. thesis, University of Massachusetts.

Tiedemann, J. (2011). *Bitext Alignment*. Number 14 in Synthesis Lectures on Human Language Technologies, Graeme Hirst (ed).

Tillmann, C. and Hewavitharana, S. (2013). A unified alignment algorithm for bilingual data. *Natural Language Engineering*, 19:33–60.

Uszkoreit, J., Ponte, J., Popat, A., and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In *Proceedings of COLING*, pages 1101–1109.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of RANLP*, pages 590–596.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, January.

Wu, D. (1994). Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of ACL*, pages 80–87.

Wu, D. (2010). Alignment. In *CRC Handbook of Natural Language Processing*, number 16, pages 367–408.

Xu, Y., Max, A., and Yvon, F. (2015). Sentence alignment for literary texts. *Linguistic Issues in Language Technology*, 12(6).

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2001). Generalized belief propagation. In *Proceedings of NIPS*, pages 689–695.

Yu, Q., Max, A., and Yvon, F. (2012). Revisiting sentence alignment algorithms for alignment visualization and evaluation. In *Proceedings of BUCC*, Istanbul, Turkey.

Yvon, F., Xu, Y., Pillias, C., Cubaud, P., and Apidianaki, M. (2016). Transread: Designing a bilingual reading experience with machine translation technologies. In *Proceedings of NAACL'16 (demo session)*.

## 8. Language Resource References

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT summit*, 5:79–86.

# Appendix

Table 8 contains the text of a passage of Jean-Jacques Rousseau's "Les Confessions", corresponding to the alignment prediction matrix in Figure 3.

| | | | |
|---|---|---|---|
| $\mathrm{en}_{10}$<br>$\mathrm{en}_{11}$ | My mother's circumstances were more affluent; she was daughter of a Mons.<br>Bernard, minister, and possessed a considerable share of modesty and beauty; indeed, my father found some difficulty in obtaining her hand. | Ma mère, fille du ministre Bernard, était plus riche: elle avait de la sagesse et de la beauté.<br>Ce n'était pas sans peine que mon père l'avait obtenue. | $\mathrm{fr}_{19}$<br>$\mathrm{fr}_{20}$ |
| $\mathrm{en}_{12}$ | The affection they entertained for each other was almost as early as their existence; at eight or nine years old they walked together every evening on the banks of the Treille, and before they were ten, could not support the idea of separation. | Leurs amours avaient commencé presque avec leur vie; dès l'âge de huit à neuf ans ils se promenaient ensemble tous les soirs sur la Treille; à dix ans ils ne pouvaient plus se quitter. | $\mathrm{fr}_{21}$ |
| $\mathrm{en}_{13}$ | A natural sympathy of soul confined those sentiments of predilection which habit at first produced; born with minds susceptible of the most exquisite sensibility and tenderness, it was only necessary to encounter similar dispositions; that moment fortunately presented itself, and each surrendered a willing heart. | La sympathie, l'accord des âmes, affermit en eux le sentiment qu'avait produit l'habitude.<br>Tous deux, nés tendres et sensibles, n'attendaient que le moment de trouver dans un autre la même disposition, ou plutôt ce moment les attendait eux-mêmes, et chacun d'eux jeta son coeur dans le premier qui s'ouvrit pour le recevoir. | $\mathrm{fr}_{22}$<br>$\mathrm{fr}_{23}$ |
| $\mathrm{en}_{14}$<br>$\mathrm{en}_{15}$ | The obstacles that opposed served only to give a decree of vivacity to their affection, and the young lover, not being able to obtain his mistress, was overwhelmed with sorrow and despair.<br>She advised him to travel – to forget her. | Le sort, qui semblait contrarier leur passion, ne fit que l'animer .<br>Le jeune amant ne pouvant obtenir sa maîtresse se consumait de douleur: elle lui conseilla de voyager pour l'oublier . | $\mathrm{fr}_{24}$<br>$\mathrm{fr}_{25}$ |
| $\mathrm{en}_{16}$ | He consented – he travelled, but returned more passionate than ever, and had the happiness to find her equally constant, equally tender. | Il voyagea sans fruit, et revint plus amoureux que jamais.<br>Il retrouva celle qu'il aimait tendre et fidèle. | $\mathrm{fr}_{26}$<br>$\mathrm{fr}_{27}$ |

Table 8: The correct alignment of a passage of Jean-Jacques Rousseau's "Les Confessions", corresponding to the alignment prediction matrix in Figure 3.