

# Minimum Bayes-Risk Word Alignments of Bilingual Texts

Shankar Kumar and William Byrne

Center for Language and Speech Processing, Johns Hopkins University,  
3400 North Charles Street, Baltimore, MD, 21218, USA  
{skumar,byrne}@jhu.edu

## Abstract

We present Minimum Bayes-Risk word alignment for machine translation. This statistical, model-based approach attempts to minimize the expected risk of alignment errors under loss functions that measure alignment quality. We describe various loss functions, including some that incorporate linguistic analysis as can be obtained from parse trees, and show that these approaches can improve alignments of the English-French Hansards.

## 1 Introduction

The automatic determination of word alignments in bilingual corpora would be useful for Natural Language Processing tasks such as statistical machine translation, automatic dictionary construction, and multilingual document retrieval. The development of techniques in all these areas would be facilitated by automatic performance metrics, and alignment and translation quality metrics have been proposed (Och and Ney, 2000b; Papineni et al., 2002). However, given the difficulty of judging translation quality, it is unlikely that a single, global metric will be found for any of these tasks. It is more likely that specialized metrics will be developed to measure specific aspects of system performance. This is even desirable, as these specialized metrics could be used in tuning systems for particular applications.

We have applied Minimum Bayes-Risk (MBR) procedures developed for automatic speech recognition (Goel and Byrne, 2000) to word alignment of bitexts. This is a modeling approach that can be used with statistical models of speech and language to develop algorithms that are optimized for specific loss functions. We will discuss loss functions that can

be used for word alignment and show how the overall alignment process can be improved by the use of loss functions that incorporate linguistic features, such as parses and part-of-speech tags.

## 2 Word-to-Word Bitext Alignment

We will study the problem of aligning an English sentence to a French sentence and we will use the word alignment of the IBM statistical translation models (Brown et al., 1993).

Let  $e = e_0^l$  and  $f = f_1^m$  denote a pair of translated English and French sentences. An English word is defined as an ordered pair  $e = (j, w) : w \in V_E, j \in \{0, 1, 2, \dots, l\}$ , where the index  $j$  refers to the position of the word in the English sentence;  $V_E$  is the vocabulary of English; and the word at position 0 is the NULL word to which “spurious” French words may be aligned. Similarly, a French word is written as  $f = (i, w) : w \in V_F, i \in \{1, 2, 3, \dots, m\}$ .

An *alignment* between  $e$  and  $f$  is defined to be a sequence  $a = [a_1, a_2, \dots, a_m]$  where  $a_i : i \in \{1, 2, \dots, m\}, a_i \in \{0, 1, \dots, l\}$ . Under the alignment  $a$ , the French word  $f_i$  is connected to the English word  $e_{a_i}$ . For every alignment  $a$ , we define a *link set*  $B$  defined as  $B = \{b_1, b_2, \dots, b_m\}$  whose elements are given by the alignment links  $b = (i, j) : j = a_i, i = 1, 2, \dots, m$ .

## 3 Alignment Loss Functions

In this section we introduce loss functions to measure the quality of automatically produced alignments. Suppose we wish to compare an automatically produced alignment  $a'$  to a reference alignment  $a$ , which we assume was produced by a competent translator. We will define various loss functions  $L(B, B')$  that measure the quality of  $a'$  relative to  $a$  through their link sets  $B'$  and  $B$ .

The desirable qualities in translation are fluency

and adequacy. We assume here that both word sequences are fluent and adequate translations but that the word and phrase correspondences are unknown. It is these correspondences that we wish to determine and evaluate automatically.

We now present two general classes of loss functions that measure alignment quality. In subsequent sections, we will give specific examples of these and show how to construct decoders that are optimized for each loss function.

### 3.1 Alignment Error

The Alignment Error Rate (AER) introduced by Och and Ney (2000b) measures the fraction of links by which the automatic alignment differs from the reference alignment. Links to the NULL word are ignored. This is done by defining modified link sets for the reference alignment  $\bar{B} = B - \{(i, a_i) : a_i = 0\}$  and the automatic alignment  $\bar{B}' = B' - \{(i, a'_i) : a'_i = 0\}$ .

The reference annotation procedure allowed the human transcribers to identify which links in  $\bar{B}$  they judged to be unambiguous. In addition to the reference alignment, this gives a set of *sure links* ( $S$ ) which is a subset of  $\bar{B}$ .

AER is defined as (Och and Ney, 2000b)

$$AER(S, B; B') = 1 - \frac{|\bar{B}' \cap S| + |\bar{B}' \cap \bar{B}|}{|\bar{B}'| + |S|}. \quad (1)$$

Since our modeling techniques require loss functions rather than error rates, we introduce the Alignment Error loss function

$$\begin{aligned} L_{AE}(B, B') &= |\bar{B}| + |\bar{B}'| - 2|\bar{B} \cap \bar{B}'| \quad (2) \\ &= |\bar{B}| + |\bar{B}'| - 2 \sum_{b \in \bar{B}} \sum_{b' \in \bar{B}'} \delta_b(b'). \end{aligned}$$

We consider error rates to be “normalized” loss functions. We also note that, unlike AER,  $L_{AE}$  does not distinguish between ambiguous and unambiguous links. However, if a decoder generates an alignment  $B'$  for which  $L_{AE}(B, B')$  is zero, the AER is also zero. Therefore if AER is the metric of interest, we will design alignment procedures to minimize  $L_{AE}$ .

### 3.2 Generalized Alignment Error

We are interested in extending the Alignment Error loss function to incorporate various linguistic

features into the measurement of alignment quality. The *Generalized Alignment Error* loss is defined as

$$L_{GAE}(B, B') = 2 \sum_{b \in B} \sum_{b' \in B'} \delta_i(i') d_{ijj'}. \quad (3)$$

where  $b = (i, j)$ ,  $b' = (i', j')$  and

$$d_{ijj'} = D((j, e_j), (j', e_{j'}); f_i). \quad (4)$$

Here we have introduced the word-to-word distance measure  $D((j, e_j), (j', e_{j'}); f_i)$  which compares the links  $(i, j)$  and  $(i, j')$  as a function of the words in the translation.  $L_{GAE}$  refers to all loss functions that have the form of Equation 3. Specific loss functions are determined through the choice of  $D$ . To see the value in this, suppose  $f_i$  is a verb in the French sentence and that it is aligned in the reference alignment to  $e_j$ , the verb in the English sentence. If our goal is to ensure verb alignment, then  $D$  can be constructed to penalize any link  $(e_{j'}, f_i)$  in the automatic alignment in which  $e_{j'}$  is not a verb. We will later give examples of distances in which  $L_{GAE}$  is based on Part-of-Speech (POS) tags, parse tree distances, and automatically determined word clusters. We note that the  $L_{GAE}$  can almost be reduced to  $L_{AE}$ , except for the treatment of NULL in the English sentence.

## 4 Minimum Bayes-Risk Decoding For Automatic Word Alignment

We present the Minimum Bayes-Risk alignment formulation and derive MBR alignment procedures under the loss functions of Section 3.

Given a translated pair of English-French sentences  $(e, f)$ , the decoder  $\delta(e, f)$  produces an alignment  $B' = \delta(e, f)$ . Relative to a reference alignment  $B$ , the decoder performance is measured as  $L(B, \delta(e, f))$ . Our goal is to find the decoder that has the best performance over all translated sentences. This is measured through the Bayes Risk  $R(\delta(e, f)) = E_{P(B|f,e)}[L(B, \delta(e, f))]$ . The expectation is taken with respect to the true distribution  $P(B|f, e)$  that describes “human quality” alignments of translations as they are found in bitext.

Given a loss function and a probability distribution, it is well known that the decision rule which minimizes the Bayes Risk is given by the following expression (Bickel and Doksum, 1977; Goel and

Byrne, 2000).

$$\hat{B} = \operatorname{argmin}_{B' \in \mathcal{B}} \sum_{B \in \mathcal{B}} L(B, B') P(B|f, e). \quad (5)$$

Several modeling assumptions have been made to obtain this form of the decoder. We do not have access to the true distribution over translations. We therefore use statistical MT models to approximate  $P(B|f, e)$ . We furthermore assume that the space of alignment alternatives can be restricted to an *alignment lattice*  $\mathcal{B}$ , which is a compact representation of the most likely word alignments of the sentence pair  $(e, f)$  under the baseline models.

It is clear from Equation 5 that the MBR decoder is determined by the loss function. The *Sentence Alignment Error* refers to the loss function that gives a penalty of 1 for any errorful alignment:  $L_{0/1}(B, B') = 1 - 1_{B'}(B)$ , where  $1_{B'}$  is the indicator function of the set  $B'$ . The MBR decoder under this loss can easily be seen to be the Maximum Likelihood (ML) alignment under the MT models:  $\hat{B} = \operatorname{argmax}_{B'} P(B'|f, e)$ . This illustrates why we are interested in MBR decoders based on other loss functions: the ML decoder is optimal with respect to a loss function that is overly harsh. It does not distinguish between different types of alignment errors and good alignments receive the same penalty as poor alignments. Moreover, such a harsh penalty is particularly inappropriate when unambiguous word-to-word alignments cannot be provided in all cases even by human translators who produce the reference alignments. The AER makes an explicit distinction between ambiguous and unambiguous word alignments. Ideally, the decoder should be able to do so as well. Motivated by this, the MBR hypothesis can be thought of as the *consensus hypothesis* under a particular loss function: Equation 5 selects the hypothesis that is, in an average sense, close to the other likely hypotheses. In this way, ambiguity can be reduced by selecting the hypothesis that is “most similar” to the collection of most likely competing hypotheses.

We now describe the alignment lattice (Section 4.1) and introduce the lattice based probabilities required for the MBR alignment (Section 4.2). The derivation of the MBR alignment under the AE and GAE loss functions is presented in Sections 4.3 and 4.4.

#### 4.1 Alignment Lattice $\mathcal{B}$

The lattice  $\mathcal{B}$  is represented as a Weighted Finite State Transducer (WFST) (Mohri et al., 2000)  $\mathcal{B} = (Q, \Lambda, \kappa, F, \mathcal{T})$  with a finite set of states  $Q$ , a set of transition labels  $\Lambda$ , an initial state  $\kappa$ , the set of final states  $F$ , and a finite set of transitions  $\mathcal{T}$ . A transition in this WFST is given by  $t = (p, q, b, s)$  where  $p$  is the starting state,  $q$  is the ending state,  $b \in \Lambda$  is the alignment link and  $s$  is the weight. For an English sentence of length  $l$  and a French sentence of length  $m$ , we define  $\Lambda$  as  $\Lambda = \{(i, j) : i \in \{1, 2, \dots, m\}, j \in \{0, 1, \dots, l\}\}$ .

A *complete path* through the WFST is a sequence of transitions given by  $T = \{(p_k, q_k, b_k, s_k)\}_{k=1}^n$  such that  $p_1 = \kappa$  and  $q_n \in F$ . Each complete path defines an alignment link set  $B = \{b_k\}_{k=1}^n$ . When we write  $B \in \mathcal{B}$ , we mean that  $B$  is derived from a complete path through  $\mathcal{B}$ . This allows us to use alignment models in which the probability of an alignment can be written as a sum over alignment link weights, i.e.  $\log P(B, f|e) = \sum_{k=1}^n s_k$ .

#### 4.2 Alignment Link Posterior Probability

We first introduce the *lattice transition posterior probability* of each transition  $t = (p, q, b, s)$  in the lattice

$$P(t|f, e) = \sum_{B \in \mathcal{B}} \psi_B(t) P(B|f, e) \quad (6)$$

where  $\psi_B(t)$  is 1 if  $b \in B$  and 0 otherwise. The lattice transition posterior probability is the sum of the posterior probabilities of all lattice paths passing through the transition  $t$ . This can be computed very efficiently with a forward-backward algorithm on the alignment lattice (Wessel et al., 1998).  $P(B|f, e)$  is the posterior probability of an alignment link set which can be written as

$$P(B|f, e) = \frac{P(B, f|e)}{\sum_{B' \in \mathcal{B}} P(B', f|e)}. \quad (7)$$

We now define the *alignment link posterior probability* for a link  $b = (i, j)$

$$P(b|f, e) = \sum_{t' \in \mathcal{T}} \delta_b(b') P(t'|f, e) \quad (8)$$

where  $t' = (p', q', b', s')$ . This is the probability that any two words  $(f_i, e_j)$  are aligned given all the alignments in the lattice  $\mathcal{B}$ .

### 4.3 MBR Alignment Under $L_{AE}$

In this section we derive MBR alignment under the Alignment Error loss function (Equation 2). The optimal decoder has the form (Equation 5)

$$\hat{B} = \operatorname{argmin}_{B' \in \mathcal{B}} \sum_{B \in \mathcal{B}} L_{AE}(B, B') P(B|f, e). \quad (9)$$

The summation is equal to

$$\begin{aligned} & |\bar{B}'| + \sum_{B \in \mathcal{B}} |\bar{B}| P(B|f, e) \\ & - 2 \sum_{b' \in \bar{B}'} \left\{ \sum_{B \in \mathcal{B}} \sum_{b \in \bar{B}} \delta_b(b') P(B|f, e) \right\}. \end{aligned}$$

If  $\bar{\mathcal{T}} \subseteq \mathcal{T}$  is the subset of transitions ( $t = (p, q, b, s)$ ) that do not contain links with the NULL word, we can simplify the bracketed term as

$$\begin{aligned} & \sum_{B \in \mathcal{B}} \sum_{b \in \bar{B}} \delta_b(b') P(B|f, e) \\ & = \sum_{B \in \mathcal{B}} \sum_{t \in \bar{\mathcal{T}}} \psi_B(t) \delta_b(b') P(B|f, e) \\ & = \sum_{t \in \bar{\mathcal{T}}} \delta_b(b') \sum_{B \in \mathcal{B}} \psi_B(t) P(B|f, e) \\ & = \sum_{t \in \bar{\mathcal{T}}} \delta_b(b') P(t|f, e) \end{aligned}$$

For an alignment link  $b' \in \bar{B}'$  we note that  $\sum_{t \in \bar{\mathcal{T}}} \delta_b(b') P(t|f, e) = P(b'|f, e)$ . Therefore, the MBR alignment (Equation 9) can be found in terms of the modified link weight for each alignment link  $b' = (i', j')$

$$\hat{B} = \operatorname{argmin}_{B' \in \mathcal{B}} \sum_{b' \in \bar{B}'} (1 - 2P(b'|f, e)). \quad (10)$$

We can rewrite the above equation as

$$\begin{aligned} \hat{B} & = \operatorname{argmin}_{B' \in \mathcal{B}} \sum_{b' \in B'} y_{b'} \\ y_{b'} & = \begin{cases} 1 - 2P(b'|f, e) & j' \neq 0 \\ 0 & j' = 0. \end{cases} \quad (11) \end{aligned}$$

### 4.4 MBR Alignment Under $L_{GAE}$

We now derive MBR alignment under the Generalized Alignment Error loss function (Equation 3). The optimal decoder has the form (Equation 5)

$$\hat{B} = \operatorname{argmin}_{B' \in \mathcal{B}} \sum_{B \in \mathcal{B}} L_{GAE}(B, B') P(B|f, e). \quad (12)$$

The summation can be rewritten as

$$\begin{aligned} & \sum_{B \in \mathcal{B}} L_{GAE}(B, B') P(B|f, e) \\ & = \sum_{B \in \mathcal{B}} 2 \sum_{b \in B} \sum_{b' \in B'} \delta_i(i') d_{ijj'} P(B|f, e) \\ & = 2 \sum_{b' \in B'} \left\{ \sum_{B \in \mathcal{B}} \sum_{b \in B} \delta_i(i') d_{ijj'} P(B|f, e) \right\} \end{aligned}$$

where  $b = (i, j)$  and  $b' = (i', j')$ .

We can simplify the bracketed term as

$$\begin{aligned} & \sum_{B \in \mathcal{B}} \sum_{b \in B} \delta_i(i') d_{ijj'} P(B|f, e) \\ & = \sum_{B \in \mathcal{B}} \sum_{t \in \mathcal{T}} \delta_i(i') d_{ijj'} \psi_B(t) P(B|f, e) \\ & = \sum_{t \in \mathcal{T}} \delta_i(i') d_{ijj'} \sum_{B \in \mathcal{B}} \psi_B(t) P(B|f, e) \\ & = \sum_{t \in \mathcal{T}} \delta_i(i') d_{ijj'} P(t|f, e) \end{aligned}$$

where  $t = (p, q, b, s)$  and  $b = (i, j)$ .

The MBR alignment (Equation 12) can be found in terms of the modified link weight for each alignment link  $b'$

$$\begin{aligned} \hat{B} & = \operatorname{argmin}_{B' \in \mathcal{B}} \sum_{b' \in B'} z_{b'} \\ z_{b'} & = \sum_{t \in \mathcal{T}} \delta_i(i') d_{ijj'} P(t|f, e). \end{aligned} \quad (13)$$

### 4.5 MBR Alignment Using WFST Techniques

The MBR alignment procedures under the  $L_{AE}$  and  $L_{GAE}$  loss functions begin with a WFST that contains the alignment probabilities  $P(B, f|e)$  as described in Section 4.1. To build the MBR decoder for each loss function the weights on the transitions ( $t' = (p', q', b', s')$ ) of the WFST are modified according to either Equation 11 ( $s' = y_{b'}$ ) or Equation 13 ( $s' = z_{b'}$ ). Once the weights are modified, the search procedure for the MBR alignment is the same in each case. The search is carried out using a  $O(n^3)$  shortest-path algorithm (Mohri et al., 2000).

## 5 Word Alignment Experiments

We present here examples of Generalized Alignment Error loss functions based on three types of linguistic features and show how they can be incorporated into a statistical MT system to obtain automatic alignments.

## 5.1 Syntactic Distances From Parse-Trees

Suppose a parser is available that generates a parse-tree for the English sentence. Our goal is to construct an alignment loss function that incorporates features from the parse. One way to do this is to define a graph distance

$$d_{ijj'} = g(N_{e_j}, N_{e_{j'}}). \quad (14)$$

Here  $N_{e_j}$  and  $N_{e_{j'}}$  are the parse-tree leaf nodes corresponding to the English words  $e_j$  and  $e_{j'}$ . This quantity is computed as the sum of the distances from each node to their closest common ancestor. It gives a syntactic distance between any pair of English words based on the parse-tree. This distance has been used to measure word association for information retrieval (Mittendorf and Winiwarter, 2001). It reflects how strongly the words  $e_j$  and  $e_{j'}$  are bound together by the syntactic structure of the English sentence as determined by the parser. Figure 1 shows the parse tree for an English sentence in the test data with the pairwise syntactic distances between the English words corresponding to the leaf nodes.

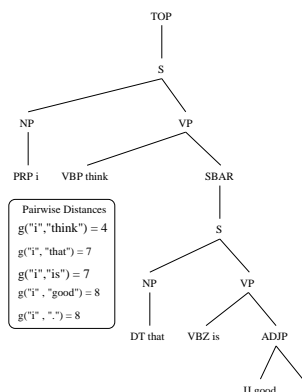


Figure 1: Parse tree for a English sentence with the pairwise syntactic distances between words.

To obtain these distances, Ratnaparkhi’s part-of-speech (POS) tagger (Ratnaparkhi, 1996) and Collins’ parser (Collins, 1999) were used to obtain parse trees for the English side of the test corpus. With  $D$  defined as in Equation 14, the Generalized Alignment Error loss function (Equation 3) is called the Parse-Tree Syntactic Distance ( $L_{PTSD}$ ).

## 5.2 Distances Derived From Part-of-Speech Labels

Suppose a Part-of-Speech(POS) tagger is available to tag each word in the English sentence. If  $\text{POS}(e_j)$  denotes the POS of the English word  $e_j$ , we can define the word-to-word distance measure  $D$  (Equation 4) as

$$d_{ijj'} = \begin{cases} 0 & \text{POS}(e_j) = \text{POS}(e_{j'}) \\ 1 & \text{otherwise.} \end{cases} \quad (15)$$

Ratnaparkhi’s POS tagger (Ratnaparkhi, 1996) was used to obtain POS tags for each word in the English sentence. With  $D$  specified by Equation 15, the Generalized Alignment Error loss function (Equation 3) is called the Part-Of-Speech Distance ( $L_{POSD}$ ).

## 5.3 Automatic Word Cluster Distances

Suppose we are working in a language for which parsers and POS taggers are not available. In this situation we might wish to construct the loss functions based on word classes determined by automatic clustering procedures. If  $C(e_j)$  specifies the word cluster for the English word  $e_j$ , then we define the distance

$$d_{ijj'} = \begin{cases} 0 & C(e_j) = C(e_{j'}) \\ 1 & \text{otherwise.} \end{cases} \quad (16)$$

In our experiments we obtained word clusters for English words using a statistical learning procedure (Kneser and Ney, 1991) where the total number of word classes is restricted to be 100. With  $D$  as defined in Equation 16, the Generalized Alignment Error loss function (Equation 3) is called the Automatic Word Class Distance ( $L_{AWCD}$ ).

## 5.4 IBM-3 Word Alignment Models

Since the true distribution over alignments is not known, we used the IBM-3 statistical translation model (Brown et al., 1993) to approximate  $P(B, f|e)$ . This model is specified through four components: Fertility probabilities for words; Fertility probabilities for NULL; Word Translation probabilities; and Distortion probabilities. We used a modified version of the IBM-3 distortion model (Knight and Al-Onaizan, 1998) in which each of the  $m!$  possible permutations of the French

sentence is equally likely. The IBM-3 models were trained on a subset of the Canadian Hansards French-English data which consisted of 50,000 parallel sentences (Och and Ney, 2000b). The vocabulary size was 18,499 for English and 24,198 for French. The GIZA++ toolkit (Och and Ney, 2000a) was used for training the IBM-3 models (as in (Och and Ney, 2000b)).

## 5.5 Word Alignment Lattice Generation

We obtained word alignments under the modified IBM-3 models using the finite state translation framework introduced by Knight and Al-Onaizan (1998). The finite state operations were carried out using the AT&T Finite State Machine Toolkit (Mohri et al., 2001; Mohri et al., 2000).

The WFST framework involves building a transducer for each constituent of the IBM-3 Alignment Models: the word fertility model  $M$ ; the NULL fertility model  $N$ ; and the word translation model  $T$  (Section 5.4). For each sentence pair we also built a finite state acceptor  $E$  that accepts the English sentence and another acceptor  $F$  which accepts all legal permutations of the French sentence. The alignment lattice  $\mathcal{B}$  for the sentence pair was then obtained by the following weighted finite state composition  $\mathcal{B} = E \circ M \circ N \circ T \circ F$ . In practice, the WFST obtained by the composition was pruned to a maximum of 10,000 states using a likelihood based pruning operation. In terms of AT&T Finite State Toolkit shell commands, these operations are given as:

```

fsmcompose E M | fsmcompose - N | \
fsmcompose - T | fsmcompose - F | \
fsmprune -n 10000 > \mathcal{B}

```

The finite state composition and pruning were performed using lazy implementations of algorithms provided in AT&T Finite State libraries (Mohri et al., 2000). This made the computation efficient because even though five WFSTs are composed into a potentially huge transducer, only a small portion of it is actually searched during the pruning used to generate the final lattice.

A heavily pruned alignment lattice  $\mathcal{B}$  for a sentence-pair from the test data is shown in Figure 2. For clarity of presentation, each alignment link  $b = (i, j)$  in the lattice is shown as an ordered

pair  $(x_{-j}) : (y_{-i})$  where  $x = e_j$  and  $y = f_i$  are the English and French words on the link. For each sentence, we also computed the lattice path with the highest probability  $P(\mathcal{B}|f, e)$ . This gives the ML alignment under the statistical MT models that will give our baseline performance under the various loss functions.

## 5.6 Performance Under The Alignment Error Rates

Our unseen test data consisted of 207 French-English sentence pairs from the Hansards corpus (Och and Ney, 2000b). These sentence pairs had at most 16 words in the French sentence; this restriction on the sentence length was necessary to control the memory requirements of the composition.

### 5.6.1 MBR Consensus Alignments

In the previous sections we introduced a total of four loss functions:  $L_{AE}$ ,  $L_{PTSD}$ ,  $L_{POSD}$  and  $L_{AWCD}$ . Using either Equation 11 or 13, an MBR decoder can be constructed for each. These decoders are called MBR-AE, MBR-PTSD, MBR-POSD, and MBR-AWCD, respectively.

### 5.6.2 Evaluation Metrics

The performance of the four decoders was measured with respect to the alignments provided by human experts (Och and Ney, 2000b). The first evaluation metric used was the Alignment Error Rate (Equation 1). We also evaluated each decoder under the *Generalized Alignment Error Rates* (GAER). These are defined as:

$$GAER(\mathcal{B}, \mathcal{B}') = \frac{L_{GAE}(\mathcal{B}, \mathcal{B}')}{|\mathcal{B}| + |\mathcal{B}'|}. \quad (17)$$

There are six variants of *GAER*. These arise when  $L_{GAE}$  is specified by  $L_{PTSD}$ ,  $L_{POSD}$  or  $L_{AWCD}$ . There are two versions of each of these: one version is sensitive only to sure (S) links. The other version considers all (A) links in the reference alignment. We therefore have the following six Generalized Alignment Error Rates: PTSD-S, POSD-S, AWCD-S, and PTSD-A, POSD-A, AWCD-A. We say we have a matched condition when the same loss function is used in both the error rate and the decoder design.

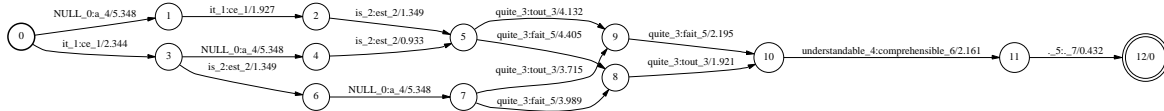


Figure 2: A heavily pruned alignment lattice for the English-French sentence pair  $e$ ="it is quite understandable ."  $f$ ="ce est tout a fait comprehensible .".

### 5.6.3 Decoder Performance

The performance of the decoders under various loss functions is given in Table 1. We observe that in none of these experiments was the ML decoder found to be optimal. In all instances, the MBR decoder tuned for each loss function was the best performing decoder under the corresponding error rate. In particular, we note that alignment performance as measured under the AER metric can be improved by using MBR instead of ML alignment. This demonstrates the value of finding decoding procedures matched to the performance criterion of interest.

We observe some affinity among the loss functions. In particular, the ML decoder performs better under the AER than any of the MBR-GAE decoders. This is because the  $L_{0/1}$  loss, for which the ML decoder is optimal, is closer to the  $L_{AE}$  loss than any of the  $L_{GAE}$  loss functions. The NULL symbol is treated quite differently under  $L_{AE}$  and  $L_{GAE}$ , and this leads to a large mismatch between the MBR-GAE decoders and the AER metric. Similarly, the performance of the MBR-POS decoder degrades significantly under the AWCD-S and AWCD-A metrics. Since there are more word clusters (100) than POS tags (55), the MBR-POS decoder is therefore incapable of producing hypotheses that can match the word clusters used in the AWCD metrics.

## 6 Discussion And Conclusions

We have presented a Minimum Bayes-Risk decoding strategy for obtaining word alignments of bilingual texts. MBR decoding is a general formulation that allows the construction of specialized decoders from general purpose models. The strategy aims at direct minimization of the expected risk of alignment errors under a given alignment loss function.

We have introduced several alignment loss functions to measure the alignment quality. These incorporate information from varied analyses, such as parse trees, POS tags, and automatically derived

word clusters. We have derived and implemented lattice based MBR consensus decoders under these loss functions. These decoders rescore the lattices produced by maximum likelihood decoding to produce the optimal MBR alignments.

We have chosen to present MBR decoding using the IBM-3 statistical MT models implemented via WFSTs. However MBR decoding is not restricted to this framework. It can be applied more broadly using other MT model architectures that might be selected for reasons of modeling fidelity or computational efficiency.

We have presented these alignment loss functions to explore how linguistic knowledge might be incorporated into machine translation systems without building detailed statistical models of these linguistic features. However we stress that the MBR decoding procedures described here do not preclude the construction of complex MT models that incorporate linguistic features. The application of such models, which could be trained using conventional maximum likelihood estimation techniques, should still benefit by the application of MBR decoding techniques.

In future work we will investigate loss functions that incorporate French and English parse-tree information into the alignment decoding process. Our ultimate goal, towards which this work is the first step, is to construct loss functions that take advantage of linguistic structures such as syntactic dependencies found through monolingual analysis of the sentences to be aligned. Recent work (Hwa et al., 2002) suggests that translational correspondence of linguistic structures can indeed be useful in projecting parses across languages. Our ideal would be to construct MBR decoders based on loss functions that are sensitive both to word alignment as well as to agreement in higher level structures such as parse trees. In this way ambiguity present in word-to-word alignments will be resolved by the alignment of linguistic structures.

		Generalized Alignment Error Rates					
Decoder	AER	PTSD-S	POSD-S	AWCD-S	PTSD-A	POSD-A	AWCD-A
ML	18.13	3.13	4.35	4.69	29.39	51.36	54.58
MBR-AE	<b>14.87</b>	1.34	1.89	1.94	19.81	36.42	38.58
MBR-PTSD	23.26	0.62	0.69	0.82	<b>14.45</b>	26.76	28.42
MBR-POSD	28.60	2.43	0.69	3.23	15.70	<b>26.28</b>	29.48
MBR-AWCD	24.71	1.00	0.95	0.86	14.92	26.83	<b>28.39</b>

Table 1: Performance (%) of the MBR decoders under the Alignment Error and Generalized Alignment Error Rates. For each metric the error rate of the matched decoder is in bold.

MBR alignment is a promising modeling framework for the detailed linguistic annotation of bilingual texts. It is a simple model rescoring formalism that improves well trained statistical models by tuning them for particular performance criteria. Ideally, it will be used to produce decoders optimized for the loss functions that actually measure the qualities that we wish to see in newly developed automatic systems.

#### Acknowledgments

We would like to thank F. J. Och of RWTH, Aachen for providing us the GIZA++ SMT toolkit, the *mkcls* toolkit to train word classes, the Hansards 50K training and test data, and the reference word alignments and AER metric software. We would also like to thank P. Resnik, R. Hwa and O. Kolak of the Univ. of Maryland for useful discussions and help with the GIZA++ setup. We thank AT&T Labs - Research for use of the FSM Toolkit. This work was supported by an ONR MURI grant N00014-01-1-0685.

#### References

- P. J. Bickel and K. A. Doksum. 1977. *Mathematical Statistics: Basic Ideas and Selected topics*. Holden-Day Inc., Oakland, CA, USA.
- P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.
- M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.
- V. Goel and W. Byrne. 2000. Minimum Bayes-risk automatic speech recognition. *Computer Speech and Language*, 14(2):115–135.
- R. Hwa, P. Resnik, A. Weinberg, and O. Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of ACL-2002*. To appear.
- R. Kneser and H. Ney. 1991. Forming word classes by statistical clustering for statistical language modelling. In *The 1st Quantitative Linguistics Conference*, Trier, Germany.
- K. Knight and Y. Al-Onaizan. 1998. Translation with finite-state devices. In *Proceedings of the AMTA Conference*, pages 421–437, Langhorne, PA, USA.
- M. Mittendorf and W. Winiwarter. 2001. Experiments with the use of syntactic analysis in information retrieval. In *Proceedings of the 6th International Workshop on Applications of Natural Language and Information Systems*, Bonn, Germany.
- M. Mohri, F. C. N. Pereira, and M. Riley. 2000. The design principles of a weighted finite-state transducer library. *Theoretical Computer Science*, 231(1):17–32.
- M. Mohri, F. Pereira, and M. Riley, 2001. *ATT General-purpose finite-state machine software tools*. <http://www.research.att.com/sw/tools/fsm/>.
- F. Och and H. Ney. 2000a. A comparison of alignment models for statistical machine translation. In *Proceedings Of 18th Conference On Computational Linguistics*, pages 1086–1090, Saarbrucken, Germany.
- F. Och and H. Ney. 2000b. Improved statistical alignment models. In *Proceedings of ACL-2000*, pages 440–447, Hong Kong, China.
- K. Papineni, S. Roukos, T. Ward, J. Henderson, and F. Reeder. 2002. Corpus-based comprehensive and diagnostic mt evaluation: Initial arabic, chinese, french, and spanish results. In *Proceedings of HLT 2002*.
- A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142, Philadelphia, PA, USA.
- F. Wessel, K. Macherey, and R. Schlueter. 1998. Using word probabilities as confidence measures. In *Proceedings of ICASSP-98*, pages 225–228, Seattle, WA, USA.