Fast and Accurate Sentence Alignment of Bilingual Corpora

Robert C. Moore

Microsoft Research
Redmond, WA 98052, USA
bobmoore@microsoft.com

Abstract. We present a new method for aligning sentences with their translations in a parallel bilingual corpus. Previous approaches have generally been based either on sentence length or word correspondences. Sentence-length-based methods are relatively fast and fairly accurate. Word-correspondence-based methods are generally more accurate but much slower, and usually depend on cognates or a bilingual lexicon. Our method adapts and combines these approaches, achieving high accuracy at a modest computational cost, and requiring no knowledge of the languages or the corpus beyond division into words and sentences.

1 Introduction

Sentence-aligned parallel bilingual corpora have proved very useful for applying machine learning to machine translation, but they usually do not originate in sentence-aligned form. This makes the task of aligning such a corpus of considerable interest, and a number of methods have been developed to solve this problem. Ideally, a sentence-alignment method should be fast, highly accurate, and require no special knowledge about the corpus or the two languages.

Kay and Röscheisen [1][2] developed an iterative relaxation approach to sentence alignment, but it was not efficient enough to apply to large corpora. The first approach shown to be effective at aligning large corpora was based on modeling the relationship between the lengths of sentences that are mutual translations. Similar algorithms based on this idea were developed independently by Brown, et al. [3] and Gale and Church [4][5]. Subsequently, Chen [6] developed a method based on optimizing word-translation probabilities which he showed gave better accuracy than the sentence-length-based approach, but was "tens of times slower than the Brown and Gale algorithms" [6, p. 15]. Wu [7] used a version of Gale and Church's method adapted to Chinese along with lexical cues in the form of a small corpus-specific bilingual lexicon to improve alignment accuracy in text regions containing multiple sentences of similar length. Melamed [8][9] also developed a method based on word correspondences, for which he reported [8] sentence-alignment accuracy slightly better than Gale and Church. Simard and Plamondon [10] developed a two-pass approach, in which a method similar to Melamed's identifies points of correspondence in the text that constrain a second-pass search that uses a statistical translation model.

All these prior methods require particular knowledge about the corpus or the languages involved. The length-based methods require no special knowledge about the

languages, but the implementations of Brown et al. and Gale and Church require either corpus-dependent anchor points, or prior alignment of paragraphs to constrain the search. The word-correspondence-based methods of Chen and Melamed do not require this sort of information about the corpus, but they either require an initial bilingual lexicon, or they depend on finding cognates in the two languages to suggest word correspondences. Wu's method also requires that the bilingual lexicon be externally supplied. Simard and Plamondon's approach relies on the existence of cognates for the first pass, and a previously-trained word-translation model for the second pass.

We have developed a hybrid sentence-alignment method, using previous sentence-length-based and word-correspondence-based models, that is fast, very accurate, and requires only that the corpus be separated into words and sentences. In a direct comparison with a length-based model that is a slight modification of Brown et al.'s, we find our hybrid method has a precision error rate 5 to 13 times smaller, and a recall error rate 5 to 38 times smaller. Moreover, the ratio of the computation times required for our method, vs. the length-only-based method, is less than 3 for easy to align material and seems to asymptotically approach 1 as the material becomes harder to align, which is when our advantage in precision and recall is greatest.

2 Description of the Algorithm

Our algorithm combines techniques adapted from previous work on sentence and word alignment in a three-step process. We first align the corpus using a modified version of Brown et al.'s sentence-length-based model. We employ a novel search-pruning technique to efficiently find the sentence pairs that align with highest probability without the use of anchor points or larger previously aligned units. Next, we use the sentence pairs assigned the highest probability of alignment to train a modified version of IBM Translation Model 1 [11]. Finally, we realign the corpus, augmenting the initial alignment model with IBM Model 1, to produce an alignment based both on sentence length and word correspondences. The final search is confined to the minimal alignment segments that were assigned a nonnegligible probability according to the initial alignment model, which reduces the size of the search space so much that this alignment is actually faster than the initial alignment, even though the model is much more expensive to apply to each segment.

Our method is simliar to Wu's [7] in that it uses both sentence length and lexical correspondences to derive the final alignment, but since the lexical correspondences are themselves derived automatically, we require no externally supplied lexicon. We discuss each of the steps of our approach in more detail below.

2.1 Sentence-Length-Based Alignment

Brown et al. [3] assume that every parallel corpus can be aligned in terms of a sequence of minimal alignment segments, which they call "beads", in which sentences align 1-to-1, 1-to-2, 2-to-1, 1-to-0, or 0-to-1.[1] The alignment model is a generative probabilistic

[1] This assumption fails occasionally when there is an alignment of 2-to-2 or 3-to-1, etc. This is of little concern, however, because it is sufficient for our purposes to extract the 1-to-1

model for predicting the lengths of the sentences composing sequences of such beads. The model assumes that each bead in the sequence is generated according to a fixed probability distribution over bead types, and for each type of bead there is a submodel that generates the lengths of the sentences composing the bead.

For the 1-to-0 and 0-to-1 bead types, there is only one sentence in each bead, and the lengths of those sentences are assumed to be distributed according a model based on the observed distribution of sentence lengths in the text in the corresponding language. For all the other beads types (1-to-1, 2-to-1, and 1-to-2), the length(s) of the sentence(s) of the first (source) language are assumed to be distributed according to the same model used in the 1-to-0 case, and the total length of the sentence(s) of the second (target) language in the bead is assumed to be distributed according to a model conditioned on the total length of the sentence(s) of the source language in the bead. Brown et al. assume that the logarithm of the ratios of the length l_t of the sentence(s) of the target language to the length l_s of the corresponding sentence(s) of the source language varies according to a Gaussian distribution with mean μ and variance σ^2,

$$P(l_t|l_s) = \alpha \; exp(-((log(l_t/l_s) - \mu)^t/2\sigma^2)) \; , \tag{1}$$

where α is chosen to make $P(l_t|l_s)$ sum to 1 for positive integer values of l_t.

The major difference between our sentence-length-based alignment model and that of Brown et al. is in how the conditional probability $P(l_t|l_s)$ is estimated. Our model assumes that l_t varies according to a Poisson distribution whose mean is simply l_s times the ratio r of the mean length of sentences of the target language to the mean length of sentences of the source language:

$$P(l_t|l_s) = exp(-l_s r)(l_s r)^{l_t}/(l_t!) \; . \tag{2}$$

The idea is that each word of the source language translates into some number of words in the target language according to a Poisson distribution, whose mean can be estimated simply as the ratio of the mean sentence lengths in the two languages. This model is simple to estimate because it has no hidden parameters, whereas at least the variance σ^2 needs to be estimated iteratively using EM in Brown et al.'s Gaussian model. Moreover, when we compared the two models on several thousand sentences of hand-aligned data, we found that the Poisson distribution actually fit the data slightly better than the best-fitting Gaussian distribution of the form used by Brown et al.

There are a few other minor differences between the two models. Brown et al. estimate marginal distributions of sentence lengths in the two languages using the raw relative frequencies in the corpus to estimate the probabilities of the lengths of shorter sentences, and smooth the estimates for the lengths of longer sentences by fitting to the tail of a Poisson distribution. In contrast, we simply use the raw relative frequencies to estimate the probability of every observed sentence length. This only affects the estimates for particularly long, and therefore rare, sentence lengths, which should have no appreciable effect on the performance of the model. We also found that the performance of the model was rather insensitive to the exact values of the probabilities assigned to

alignments, which account for the vast majority of most parallel corpora and are in practice the only alignments that are currently used for training machine translation systems.

the various bead types, so we simply chose rough values close to those reported by Brown et al. and Gale and Church, rather than tuning them by re-estimation as Brown et al. do. We experimented with initializing the model with these values and iteratively re-estimating to the optimal values for our data, but we never saw a significant difference in the output of alignment as a result of re-estimating these parameters. Finally, Brown et al. also include paragraph boundary beads in their model, which we omit, in part because paragraph boundary information was not present in our data.

Our intention in making these modifications to the model of Brown et al. is not to improve its accuracy in sentence alignment, and we certainly do not claim to have done so. In fact, we believe that the differences are so slight that the models should perform comparably. The practical difference between the two models is that because ours has no hidden parameters, we don't need to use EM or any other iterative parameter re-estimation method, which makes our variant much faster to use in practice.

Search Issues The standard approach to solving alignment problems is to use dynamic programming (DP). In an exhaustive DP alignment search, one iteratively computes some sort of cost function for all possible points of correspondence between the two sequences to be aligned. For the sentence alignment problem, the number of such points is approximately the product of the numbers of sentences in each language; so it is clearly infeasible to do an exhaustive DP search for a large corpus. The search must therefore be pruned in some way, which is the approach we have followed, as have Brown et al., Gale and Church, and Chen. Our method of pruning, however, is novel and has proved quite effective.

Notice that unless there are extended segments of one language not corresponding to anything in the other language, the true points of correspondence should all be close to proportionately the same distance from the beginning of each text. For example, the only way a point 30% of the way along the text in the source language would be likely to correspond to a point 70% of the way along the text in the target language is if there were some major insertions and/or deletions in one or both of the texts. Following Melamed, we think of the set of possible points of correspondence as forming a matrix, and the set of points closest to proportionately the same distance from the beginning of each text as "the main diagonal".

Our pruned DP search starts by doing an exhaustive search, but only in a narrow fixed-width band around the main diagonal. Unless there are extended segments of one language not corresponding to anything in the other language, the best alignment of the two texts will usually fall within this band. But how do we know whether this is the case? Our heuristic is to look at an approximate best alignment within the band, and find the point where it comes closest to one of the boundaries of the band. If the approximate best alignment never comes closer than a certain minimum distance from the boundaries of the band, we assume that the best alignment within the band is actually the best possible alignment, and the search terminates. Otherwise, we widen the band and iterate. While we have no proof that this heuristic will always work, we have never seen it commit a search error in practice. Our conjecture is that if the search band is too narrow to contain the true best alignment, the constrained best alignment will basically be a random walk in those regions where the true best alignment is excluded. If the

size of the excluded regions of the true best alignment is large, the probability of this random walk never coming close to a boundary is small.

In this phase of our algorithm, the main goal is to find all the high-probability 1-to-1 beads to use for training a word-translation model. We find these beads by performing the forward-backward probability computation, as described by Rabiner [12], using the initial search described above as the forward pass. To speed up the backward pass of this search, we start by considering only points that have survived the first pass pruning, and we further prune out (and do not extend) any of these points that receive a very low total probability in the backward pass.

2.2 Word-Translation Model

In the next phase of our algorithm, we use the highest probability 1-to-1 beads from the initial alignment to train a word-translation model. We use a threshold of 0.99 probability of correct alignment to ensure reliable training data, and in our experiments this makes use of at least 80% of the corpus. For our word-translation model, we use a modified version of the well-known IBM Translation Model 1 [11].

The general picture of how a target language sentence t is generated from a source language sentence s consisting of l words, $s_1 \ldots s_l$, in the IBM translation models is as follows: First, a length m is selected for t. Next, for each word position in t, a generating word in s (including the null word s_0) is selected. Finally, for each pair of a position in t and its generating word in s, a target language word is chosen to fill the target position. Model 1 makes the assumptions that all possible lengths for t (less than some arbitrary upper bound) have a uniform probability ϵ; all possible choices of the source language generating words are equally likely; and the probability $tr(t_j|s_i)$ of the generated target language word depends only on the generating source language word—which Brown et al. show yields

$$P(t|s) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} tr(t_j|s_i) \ . \tag{3}$$

We make two minor modifications in Model 1 for the sake of space efficiency. The translation probabilities for rare words can be omitted without much loss, since they will hardly ever be used. Therefore, to prune the size of our word-translation model, we choose a minimum number of occurrences for a word to be represented distinctly in the model, and map all words with fewer occurrences into a single token prior to computing the word-translation model. For each language, we set the threshold to be the maximum count per word that is sufficient to result in 5000 distinct words of that language, subject to an absolute minimum for the threshold of 2 occurrences per word.

In principle, Model 1 will assign a translation probability to every possible pair consisting of one of the remaining words from each language, provided the words both occur in at least one aligned sentence pair. The vast majority of these, however, will not represent true translation pairs and therefore contribute little to determining correct sentence alignment. Therefore, our second modification to Model 1 is that, in accumulating fractional counts in each iteration of EM after the first, any fractional count for a word-translation pair in a given sentence that is not greater than would be obtained by making a totally random choice is not added to the count for that translation pair.

For example, if a source language sentence contains 10 words (including the null word) and the existing model assigns to one of those words a fractional count not greater than 0.1 for generating a particular word in the target language sentence, we don't include that fractional count in the total count for that word-translation pair. To maintain the integrity of the model, we assign these fractional counts to the pair involving the null word instead. We find this reduces the size of the model by close to 90% without significantly impacting the performance of the resulting model.

We train our modified version of Model 1 by carrying out 4 iterations of EM as described by Brown, et al. [11], which we found to be an upper bound on the number of iterations needed to minimize the entropy of held out data.

2.3 Word-Correspondence-Based Alignment

For the final sentence-alignment model we use the framework of our initial sentence-length-based model, but we modify it to use IBM Model 1 in addition to the initial model. The modified model assumes that bead types and sentence lengths are generated according to the same probability distributions used by the sentence-length-based model, but we multiply the probability estimate based on these features by an estimated probability for the actual word sequences composing each bead, based on the instance of Model 1 that we have estimated from the initial alignment.

For the single sentence in a 1-to-0 or 0-to-1 bead, each word is assumed to be generated independently according to the observed relative unigram frequency f_u of the word in the text in the corresponding language. For all the other beads types (1-to-1, 2-to-1, and 1-to-2), the words of the sentence(s) of the source language are assumed to be generated according to the same model used in the 1-to-0 case; and the words of the sentence(s) of the target language in the bead are assumed to be generated depending on the words of the source language, according to the instance of Model 1 that we have estimated from the initial alignment of the corpus. In applying Model 1, we omit the factor corresponding to the assumption of uniform distribution of target sentence lengths, since we have already accounted for sentence length by incorporating our original alignment model. For example, if s is a source sentence of length l, t is a target sentence of length m, and $P_{1-1}(l, m)$ is the probability assigned by the initial model to a sentence of length l aligning 1-to-1 with a sentence of length m, then our combined model will estimate the probability of a 1-to-1 bead consisting of s and t as

$$P(s,t) = \frac{P_{1-1}(l,m)}{(l+1)^m}\left(\prod_{j=1}^{m}\sum_{i=0}^{l} tr(t_j|s_i)\right)\left(\prod_{i=1}^{l} f_u(s_i)\right) . \tag{4}$$

Simard and Plamondon [10] also base their second pass on IBM Model 1. However, because they essentially use *only* Model 1—without embedding it in a more general framework, as we do—they have no way to assign probabilities to 1-to-0 and 0-to-1 beads. Hence their model has no way to accomodate deletions or insertions, which they conjecture results in the low precision they observe on many of their test corpora [10, p. 77].

Since our hybrid alignment model incorporating IBM Model 1 is much more expensive to apply to a bead than our original sentence-length-based model, if we were

Table 1. Results for Manual 1 data

Alignment Method	Probability Threshold	Number Right	Number Wrong	Number Omitted	Precision Error	Recall Error
Hand-Aligned	NA	9842	1	6	0.010%	0.061%
Length Only	0.5	9832	28	16	0.284%	0.162%
Length+Words	0.5	9846	5	2	0.051%	0.020%
Length+Words	0.9	9839	3	9	0.030%	0.091%

Table 2. Results for Manual 2 data

Alignment Method	Probability Threshold	Number Right	Number Wrong	Number Omitted	Precision Error	Recall Error
Hand-Aligned	NA	17276	5	99	0.029%	0.570%
Length Only	0.5	17304	18	71	0.104%	0.409%
Length+Words	0.5	17361	2	14	0.012%	0.081%
Length+Words	0.9	17316	1	59	0.006%	0.340%

to start the alignment search over from scratch, generating the final alignment would be very slow. We limit the search, however, to the set of possible points of correspondence receiving nonnegligible probability estimates in the initial sentence-length-based alignment. Since these are only a small fraction (on the order of 10% or less) of all the possible points correspondence explored in the initial alignment search, this greatly speeds up the final alignment search. In practice, the final alignment search takes less time than the initial alignment search—far less in some cases.

3 Results

We have evaluated our method on data from two English-language computer software manuals and their Spanish translations, for which we were able to obtain hand alignments of 1-to-1 beads for comparison. The automatic and hand alignments were in close enough agreement that we were able to have all the differences examined by a fluent bilingual. In some cases we found the hand alignment to be in error and the automatic alignment to be correct. For the purposes of our analysis we assume that every alignment pair that the automatic and hand alignments agree on is correct, and that all the correct alignment pairs are found either by the hand alignment or automatic alignment.

Our evaluation metrics are precision error and recall error for 1-to-1 sentence alignments. We follow Brown et al. [3, pp. 175–176] in using precision error (which they simply call "error") on 1-to-1 beads (which they call "*ef*-beads") as an evaluation metric. Because we have complete hand alignments for the 1-to-1 beads for all our test data, however, we are also able to measure recall error, which many previous studies have not been able to estimate.

Our results on data from Manual 1 are shown in Table 1, and results from Manual 2 are shown in Table 2. For each manual, we compare results for four different alignments: hand alignment, alignment based on sentence length only at the 0.5 probability

threshold, alignment based on sentence length and word correspondence at the 0.5 probability threshold, and alignment based on sentence length and word correspondence at the 0.9 probability threshold. The probability threshold refers to a cut-off based on the probability assigned to an alignment by application of the forward-backward probability computation as discussed in Section 2.1. Since we are able to estimate this probability, rather than simply computing the most probable overall alignment sequence, we can tune the precision/recall trade-off depending on where we decide to set our threshold for accepting an alignment pair.

Examining the results in Tables 1 and 2 shows that both the precision and recall error rates for all alignments are well under 1.0%, but that recall error and precision error are considerably lower for our hybrid model than for the alignment based only on sentence length. At the probability threshold of 0.5, for Manual 1 the precision error was 5.6 times lower and the recall error was 8.0 times lower for the hybrid method, and for Manual 2 the precision error was 9.0 times lower and the recall error was 5.1 times lower for the hybrid method. For Manual 1 the precision and recall error for the hybrid method (at either the 0.5 or 0.9 probability threshold) were almost as good as on the hand-aligned data, and for Manual 2 the error rates were actually better for data aligned by the hybrid method than for the hand-aligned data.

We believe that the data we have used in these experiments is representative of much of the sort of parallel data one might encounter as training data for machine translation. However, it turns out to be fairly easy data to align, as indicated by the low error rates of both forms of automatic alignment that we applied, and by the fact that the highest probability initial alignments deviated from the main diagonal by at most 6 positions in the case of the data from Manual 1 and at most 13 positions in the case of the data from Manual 2. To test how well the algorithms perform on more difficult data, we applied both the method based only on sentence length and the hybrid method to versions of the Manual 1 data, from which single blocks of 50, 100, and 300 sentences had been deleted from one side of the corpus at a randomly chosen point.

The results of this experiment are shown in Table 3, for the 0.5 probability threshold. Examining these results shows that as the size of the deletion increases, the precision and recall error rates for the alignment based only on sentence length also increase, but the error rates for hybrid method remain essentially constant. The advantage of the hybrid method thus increases to the point that, on the data with 300 sentences deleted, the precision error is 13.0 times lower and the recall error is 37.4 times lower than with the sentence-length-only-based method.

These substantial deletions stress the search strategy as well as the alignment models, since they force the initial search to examine a much wider band around the main diagonal to find the optimal alignment. We show the effect on the total time to compute the alignments in Table 4. Of necessity, the forward pass time of the sentence-length-only-based alignment increases at least in proportion to the maximum deviation of the best alignment from the main diagonal. If the width of the search band is doubled on every iteration, then the total search time should be no more than twice the time of the last iteration, and the width of the search band should be no more than twice the maximum deviation of the best alignment from the main diagonal. This means it should be possible to carry out the iterative first pass search in time proportional to the length of

Table 3. Results for Manual 1 data with deletions

Sentences Deleted	Alignment Method	Number Right	Number Wrong	Number Omitted	Precision Error	Recall Error
0	Length Only	9832	28	16	0.284%	0.162%
50	Length Only	9761	30	39	0.306%	0.398%
100	Length Only	9677	30	73	0.309%	0.749%
300	Length Only	9368	52	187	0.552%	1.967%
0	Length+Words	9846	5	2	0.051%	0.020%
50	Length+Words	9796	6	4	0.061%	0.041%
100	Length+Words	9747	5	3	0.051%	0.031%
300	Length+Words	9550	4	5	0.042%	0.052%

Table 4. Alignment time (in seconds) for deletion experiments

Sentences Deleted	First Pass Iterations	Length Align Time	Model 1 Train Time	Length+Words Align Time	total Total
0	1	161	131	155	447
50	3	686	133	195	1013
100	5	1884	128	281	2293
300	7	4360	125	555	5040

the corpus times the size of the maximum deviation of the best alignment from the main diagonal. This seems roughly consistent with the increasing times for sentence-length-only-based alignment as the number of sentences deleted goes from 50 to 300.

Naturally, the time to train IBM Model 1 is essentially independent of the difficulty of the initial alignment. What is particularly striking, however, is that the time to perform the final alignment goes up much more slowly than the time to perform the initial alignment, due to its restriction to evaluation of points of alignment receiving nonnegligible probability in the initial alignment. As the difficulty of the alignment task increases in these experiments, the ratio of the time to perform the complete alignment process to the time to perform the initial alignment decreases from 2.8 to 1.2, with every indication that it should asymptotically approach 1.0. Thus for difficult alignment tasks, we gain the error reduction of the hybrid method at almost no additional relative cost.

4 Conclusions

It was perhaps first shown by Chen [6] that word-correspondence-based models can be used to produce higher-accuracy sentence alignment than sentence-length-based models alone. The main contribution of this work is to show how get the benefit of those higher accuracy models with only a modest additional computational cost, and without the use of anchor points, cognates, a bilingual lexicon—or any other knowledge about the corpus other than its division into words and sentences. In accomplishing this, we have made the following novel contributions to the statistical models and the search strategies used:

1. Modification of Brown et al.'s [3] sentence-length-based model to use Poisson distributions, rather than Gaussians, so that no hidden parameters need to be iteratively re-estimated.
2. A novel iterative-widening search method for alignment problems, based on detecting when the current best alignment comes near the edge of the search band, which eliminates the need for anchor points.
3. Modification of IBM Translation Model 1, eliminating rare words and low probability translations to reduce the size of the model by 90% or more.
4. Use of the probabilities computed by a relatively cheap initial model (the sentence-length-based model) to dramatically reduce the search space explored by a second more accurate, but more expensive model (the word-correspondence-based model). While this idea has been used in such fields as speech-recognition and parsing, it seems not to have been used before in bilingual alignment.

References

1. Kay, M., Röscheisen, M.: Text-Translation Alignment. Technical Report, Xerox Palo Alto Research Center (1988)
2. Kay, M., Röscheisen, M.: Text-Translation Alignment. Computational Linguistics **19(1)** (1993) 121–142
3. Brown, P.F., Lai, J.C., Mercer, R.L.: Aligning Sentences in Parallel Corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California (1991) 169–176
4. Gale, W.A., Church, K.W.: A program for Aligning Sentences in Bilingual Corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California (1991) 177–184
5. Gale, W.A., Church, K.W.: A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics **19(1)** (1993) 75–102
6. Chen, S.F.: 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio (1993) 9–16
7. Wu, D.: Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico (1994) 80–87
8. Melamed, I.D.: A Geometric Approach to Mapping Bitext Correspondence. IRCS Technical Report 96-22, University of Pennsylvania (1996)
9. Melamed, I.D.: A Portable Algorithm for Mapping Bitext Correspondence. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain (1997) 305–312
10. Simard, M., Plamondon, P.: Bilingual Sentence Alignment: Balancing Robustness and Accuracy. Machine Translation **13(1)** (1998) 59–80
11. Brown, P.F., Della Pietra, S. A., Della Pietra, V. J., Mercer, R.L.: The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics **19(2)** (1993) 263–311
12. Rabiner, L. R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE **77(2)** (1989) 257–286