Maximizing Component Quality in Bilingual Word-Aligned Segmentations

Spyros Martzoukos **Christophe Costa Florêncio** **Christof Monz**
Intelligent Systems Lab Amsterdam, University of Amsterdam
Science Park 904, 1098 XH Amsterdam, The Netherlands
{S.Martzoukos, C.CostaFlorencio, C.Monz}@uva.nl

Abstract

Given a pair of source and target language sentences which are translations of each other with known word alignments between them, we extract bilingual phrase-level segmentations of such a pair. This is done by identifying two appropriate measures that assess the quality of phrase segments, one on the monolingual level for both language sides, and one on the bilingual level. The monolingual measure is based on the notion of partition refinements and the bilingual measure is based on structural properties of the graph that represents phrase segments and word alignments. These two measures are incorporated in a basic adaptation of the Cross-Entropy method for the purpose of extracting an N-best list of bilingual phrase-level segmentations. A straight-forward application of such lists in Statistical Machine Translation (SMT) yields a conservative phrase pair extraction method that reduces phrase-table sizes by 90% with insignificant loss in translation quality.

1 Introduction

Given a pair of source and target language sentences which are translations of each other with known word alignments between them, the problem of extracting high quality bilingual phrase segmentations is defined as follows: Maximize the quality of phrase segments, i.e., groupings of consecutive words, in both language sides, subject to constraints imposed by the underlying word alignments. The purpose of this work is to provide a solution to this maximization problem and investigate the effect of the resulting high quality bilingual phrase segments on SMT. For brevity, 'phrase-level sentence segmentation' and 'phrase segment' will henceforth be simply referred to as 'segmentation' and 'segment' respectively.

The exact definition of segments' quality depends on the application. Our notion of a segmentation of maximum quality is defined as the set of consecutive words of the sentence that captures maximum collocational and/or grammatical characteristics. This implies that a sequence of tokens is identified as a segment if its fully compositional expressive power is higher than the expressive power of any combination of partial compositions. Since this definition is fairly general it is thus suitable for most NLP tasks. In particular, it is tailored to the type of segments that are suitable for the purposes of SMT and is in line with previous work (Blackwood et al., 2008; Paul et al., 2010).

With this definition in mind, we introduce a monolingual segment quality measure that is based on assessing the cost of converting one segmentation into another by means of an elementary operation. This operation, namely the 'splitting' of a segment into two segments, together with all possible segmentations of a sentence are known to form a partially ordered set (Guo, 1997). Such a construction is known as partition refinement and gives rise to the desired monolingual *surface* quality measure.

The presence of word alignments between the sentence pair provides additional structure which should not be ignored. In the language of graph theory, a segment can also be viewed as a chain, i.e., a graph in which vertices are the segment's words and

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 30–38,
Gothenburg, Sweden, April 26-30 2014.

an edge between two words exists if and only if these words are consecutive. Then, a bilingual segmentation is represented by the graph that is formed by all its source and target language chains together with edges induced by word alignments. Motivated by the phrase pair extraction methods of SMT (Och et al., 1999; Koehn et al., 2003), we focus on the connected components, or simply components of such a representation. We explain that the extent to which we can delete word alignments from a component without violating its component status, gives rise to a bilingual, purely *structural* quality measure.

The surface and structural measures are incorporated in one algorithm that extracts an N-best list of bilingual word-aligned segmentations. This algorithm, which is an adaptation of the Cross-Entropy method (Rubinstein, 1997), performs joint maximization of surface (in both languages) and structural quality measures. Components of graph representations of the resulting N-best lists give rise to high quality translation units. These units, which form a small subset of all possible (continuous) consistent phrase pairs, are used to construct SMT models. Results on Czech–English and German–English datasets show a 90% reduction in phrase-table sizes with insignificant loss in translation quality which are in line with other pruning techniques in SMT (Johnson et al., 2007; Zens et al., 2012).

2 Monolingual Surface Quality Measure

Given a sentence $s_1 s_2 ... s_k$ that consists of words s_i, $1 \leq i \leq k$, we introduce an empirical count-based measure that assesses the quality of its segmentations. By fixing a segmentation σ, we are interested in assessing the cost of perturbing σ and generating another segmentation σ'. A perturbation of σ is achieved by splitting a segment of σ into two new segments, while keeping all other segments fixed. For example, for a sentence with five words, if $\sigma : (s_1 s_2)(s_3 s_4 s_5)$, where brackets are used to distinguish the segments $s_1 s_2$ and $s_3 s_4 s_5$, then σ can be perturbed in three different ways:

- $\sigma' : (s_1)(s_2)(s_3 s_4 s_5)$, by splitting the first segment of σ.
- $\sigma'' : (s_1 s_2)(s_3)(s_4 s_5)$, by splitting at the first position of the second segment of σ.
- $\sigma''' : (s_1 s_2)(s_3 s_4)(s_5)$, by splitting at the second position of the second segment of σ,

so that σ', σ'' and σ''' are the perturbations of σ. Such perturbations are known as partition refinements in the literature (Stanley, 1997). The set of all segmentations of a sentence, equipped with the splitting operation forms a partially ordered set (Guo, 1997), and its visual representation is known as the *Hasse diagram*. Figure 1 shows such a partially ordered set for a sentence with four words.

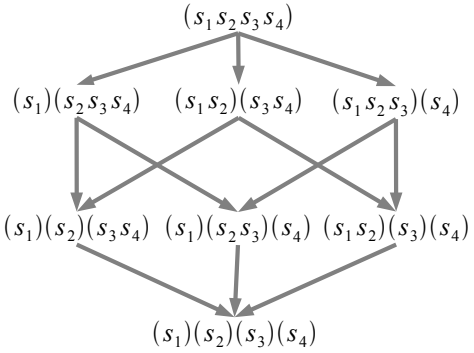

Figure 1: Hasse diagram of segmentation refinements for a sentence with four words.

The cost of perturbing a segmentation into another, i.e., the weight of a directed edge in the Hasse diagram, is calculated from n-gram counts that are extracted from a monolingual training corpus. Let $n(s)$ be the empirical count of phrase s in the corpus. Given a segmentation σ of a sentence, let $seg(\sigma)$ denote the set of σ's segments. In the above example we have for instance $seg(\sigma'') = \{s_1 s_2, s_3, s_4 s_5\}$. The probability of s in σ is given by relative frequencies

$$p_\sigma(s) = \frac{n(s)}{\sum_{s' \in seg(\sigma)} n(s')}. \quad (1)$$

The cost of perturbing σ into σ' by splitting a segment $s\bar{s}$ of σ into segments s and $\bar{s}$ is defined by

$$\text{cost}_{\sigma \to \sigma'}(s, \bar{s}) = \log \frac{p_\sigma(s\bar{s})}{p_{\sigma'}(s) p_{\sigma'}(\bar{s})}, \quad (2)$$

and we say that s and $\bar{s}$ are co-responsible for the perturbation $\sigma \to \sigma'$. Intuitively, this cost function yields the amount of energy (log of probability) that is lost when performing a perturbation. On a more

technical level, it is closely related to metric spaces on partially ordered sets (Monjardet, 1981; Orum and Joslyn, 2009), but we do not go into further details here.

The cost function admits a measure for the segments that are co-responsible for perturbing σ into σ' and we define the gain of s from the perturbation $\sigma \rightarrow \sigma'$ as

$$gain_{\sigma\rightarrow\sigma'}(s) = -\text{cost}_{\sigma\rightarrow\sigma'}(s, \bar{s}). \quad (3)$$

A segment s may be co-responsible for different perturbations, and we have to consider all such perturbations. Let

$$R(s) = \{\sigma \rightarrow \sigma' : s \notin seg(\sigma), s \in seg(\sigma')\} \quad (4)$$

denote the set of perturbations for which s is co-responsible. Then, the average gain of s in the sentence is given by

$$gain(s) = \frac{1}{|R(s)|} \sum_{\{\sigma\rightarrow\sigma'\}\in R(s)} gain_{\sigma\rightarrow\sigma'}(s). \quad (5)$$

Intuitively, $gain(s)$ measures how difficult it is to break phrase s into sub-phrases. Finally, the surface quality measure of a segmentation σ of a sentence is given by

$$g(\sigma) = \sum_{s\in seg(\sigma)} gain(s). \quad (6)$$

Note that g is a real number. The relation $g(\sigma) > g(\sigma')$ implies that σ is a better segmentation than σ'.

We conclude this section with two remarks: (i) The exact computation of $gain(s)$ for each possible segment s is computationally expensive since all perturbations need to be considered. In practice we can simply generate a random sample of no more than 1500 segmentations and compute $gain(\cdot)$ based on that sample only. (ii) Each sentence of the monolingual training corpus (from which the n-gram counts are extracted) should have the beginning and end-of-sentence tokens. The count for each of them is equal to the number of sentences in the corpus, and they are treated as regular words. Without going into further details they provide the purpose of normalization.

3 Bilingual Structural Quality Measure

Given a word-aligned sentence pair, we introduce a purely structural measure that assesses the quality of its bilingual segmentations. By 'purely structural' we mean that the focus is entirely on combinatorial aspects of the bilingual segmentations and the word alignments. For that reason we turn to a graph theoretic framework.

A segment can also be viewed as a chain, i.e., a graph in which vertices are the segment's words and an edge between two words exists if and only if these words are consecutive. Then, a source segmentation σ and a target segmentation τ are graphs that consist of source chains and target chains respectively. The graph formed by σ, τ and the translation edges induced by word alignments is thus a graph representation of a bilingual word-aligned segmentation.

We focus on a particular type of subgraphs of this representation, namely its connected components, or simply components. A component is a graph such that (a) there exists a path between any two of its vertices, and (b) there does not exist a path between a vertex of the component and a vertex outside the component. Condition (a) means, both technically and intuitively, that a component is connected and Condition (b) requires connectivity to be maximal.

Components play a key role in SMT. The most widely used strategy for extracting high quality phrase-level translations without linguistic information, namely the consistency method (Och et al., 1999; Koehn et al., 2003) is entirely based on components of word aligned unsegmented sentence pairs (Martzoukos et al., 2013). In particular, each extracted translation is either a component or the union of components. Since an unsegmented sentence pair is just one possible configuration of all possible bilingual segmentations, we consequently have no direct reason to investigate further than components.

In order to get an intuition of the measure that will be introduced in this section, we begin with an example. Figure 2, shows two different configurations of the pair (σ, τ) for the same sentence pair with known and fixed word alignments. Both configurations have the same number of edges that connect source vertices (3) and the same number of edges that connect target vertices (2). However, one would

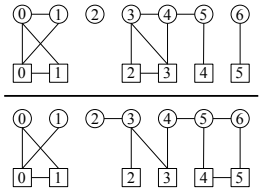

Figure 2: Graph representations of two bilingual segmentations with fixed word alignments. Source and target vertices are shown with circles and squares respectively.

expect the top configuration to represent a better bilingual segmentation. This is because it has more components (4 opposed to 2 for the bottom configuration) and because it consists of 'tighter' clusters, i.e., 'tighter' components.

A general measure that would capture this observation requires a balance between the number of edges of source and target chains, the number of components and the number of translation edges, all coupled with how these edges and vertices are connected. This might seem as a daunting task that can be tackled with a combination of heuristics, but there is actually a graph-theoretic measure that can fully describe the sought structure. We proceed with introducing this measure.

Let C denote the set of components of the graph representation of a bilingual word-aligned segmentation. We are interested in measuring the extent to which we can delete translation edges from $c \in C$, while retaining its component status. Let a_c denote the subset of translation edges that are restricted to the component c. We define the positive integer

$$\begin{aligned} gain(c) = &\text{ number of ways of} \\ &\text{deleting translation edges from } a_c, \\ &\text{while keeping } c \text{ connected,} \end{aligned} \quad (7)$$

where the option of deleting nothing is counted. Intuitively, by keeping the edges of the chains fixed the quantity $gain(c)$ measures how difficult it is to perturb a component from its connected state to a disconnected state.

Figure 3 shows two components c and c' that satisfy $gain(c) = gain(c') = 3$. Both components are equally difficult to be perturbed into a disconnected state, but only superficially. The actual structural quality of c is revealed when it is 'compared' to component $\tilde{c}$ that consists of the same source and target vertices, the same translation edges but its source vertices form exactly one chain and similarly for its target vertices; $\tilde{c}$ is essentially the 'upper bound' of c. In general, the maximum value of $gain(c)$, with

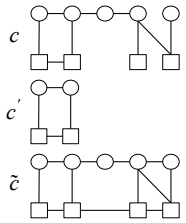

Figure 3: Superficially similar components c and c'. Comparing c with $\tilde{c}$ yields c's true structural quality.

respect to a fixed set of source and target vertices and translation edges, is attained when it consists of exactly one source chain and exactly one target chain. It is not difficult to see that the desired maximum value is always $2^{|a_c|} - 1$. In the example of Figure 3, the structural quality of c and c' is thus $3/(2^5 - 1) = 9.7\%$ and $3/(2^2 - 1) = 100\%$ respectively. Hence, the measure that evaluates the structural quality of a bilingual word-aligned segmentation (σ, τ) is given by

$$f(\sigma, \tau) = \left(\prod_{c \in C} \frac{gain(c)}{2^{|a_c|} - 1} \right)^{\frac{1}{|C|}}, \quad (8)$$

which takes values in $(0, 1]$. The relation $f(\sigma, \tau) > f(\sigma', \tau')$ implies that (σ, τ) is a better bilingual segmentation than (σ', τ').

We conclude this section with two remarks: (i) A component with no translation edges, i.e., a source or target segment whose words are all unaligned, has a contribution of $1/0$ in (8). In practice we exclude such components from C. (ii) In graph theory the quantity $gain(c)$ is known as the number of connected spanning subgraphs (CSSGs) of graph c and is the key quantity of network reliability (Valiant, 1979; Coulbourn, 1987). Finding the number of CSSGs of a general graph is a known #P-hard problem (Welsh, 1997). In our setting, graphs have specific formation (source and target chains connected via translation edges) and we are interested in the deletion of translation edges only; it is possible to

compute $gain(\cdot)$ in polynomial time, but we do not go into further details here.

4 Extracting Bilingual Segmentations with the Cross-Entropy Method

Equipped with the measures of Sections 2 and 3 we turn to extracting an N-best list of bilingual segmentations for a given sentence pair. The search space is exponential in the total number of words of the sentence pair. We propose a new approach for this task, by noting a direct connection with the combinatorial problems that can be solved efficiently and effectively with the Cross-Entropy (CE) method (Rubinstein, 1997).

The CE method is an iterative self-tuning sampling method that has applications in various combinatorial and continuous global optimization problems as well as in rare event detection. A detailed account on the CE method is beyond the scope of this work, and we thus simply describe its application to our problem.

In particular, we first establish the connection between the most basic form of the CE method and the problem of finding the best *monolingual* segmentation of a sentence, with respect to some scoring function (not necessarily the one that was introduced in Section 2). This connection yields a simple, efficient and effective algorithm for the monolingual maximization problem. Then, the transition to the bilingual level is done by incorporating the measure of Section 3 in the algorithm, thus performing joint maximization of surface and structural quality. Finally, the generation of the N-best list will be trivial.

A segmentation of a given sentence has a bit-string representation in the following way: If two consecutive words in the sentence belong to the same segment in the segmentation, then this pair of words is encoded by '1', otherwise by '0'. Such a representation is bijective and, thus, for the rest of this section, we do not distinguish between a segmentation and its bit-string representation. In this setting, the CE method takes its most basic form (De Boer et al., 2005). In a nutshell, it is a repeated application of (a) sampling bit-strings from a parametrized probability mass function, (b) scoring them and keeping only a small high-performing subsample, and (c) updating the parameters of the probability mass function based on that subsample only.

We assume no prior knowledge on the quality of bit-strings, so that they are all equally likely. In other words, each position of a randomly chosen bit-string can be either a '0' or a '1' with probability $1/2$. The aim is to tune these position probabilities towards the best bit-string, with respect to some scoring function g. In particular, let the sentence have n words and let $\ell = n - 1$ be the length of bit-strings. A bit-string labeled by an integer i is denoted by b_i and its jth bit by b_{ij}. The algorithm is as follows:

0. Initialize the bit-string position probabilities $p^0 = (p_1^0, ..., p_\ell^0) = (1/2, ..., 1/2)$ and set $M = 20\ell$ (sample size), $\rho = \lceil 1\%M \rceil$ (keep top 1% of samples), $\alpha = 0.7$ (smoothing parameter) and $t = 1$ (iteration).

1. Generate a sample $b_1, ..., b_M$ of bit-strings, each of length ℓ, such that $b_{ij} \sim$Bernoulli(p_j^{t-1}), for all $i = 1, ..., M$ and $j = 1, ..., \ell$.

1.1 Compute scores $g(b_1), ..., g(b_M)$.

1.2 Order them descendingly as $g(b_{\pi(1)}) > ... > g(b_{\pi(M)})$.

2. Focus on the best performing ones: Compute $\gamma_t = g(b_{\pi(\rho)})$; samples performing less than this threshold will be ignored.

3. Use the best performing sub-sample of $b_1, ..., b_M$ to update position probabilities:

$$p_j^t = \frac{\sum_{i=1}^M I_i(\gamma_t) b_{ij}}{\sum_{i=1}^M I_i(\gamma_t)}, \quad j = 1, ..., \ell, \tag{9}$$

where the choice function I_i is given by

$$I_i(\gamma_t) = \begin{cases} 1, & \text{if } g(b_i) > \gamma_t \\ 0, & \text{otherwise.} \end{cases}$$

4. Smooth the updated position probabilities as

$$p_j^t := \alpha p_j^t + (1 - \alpha) p_j^{t-1}, \quad j = 1, ..., \ell. \tag{10}$$

E. If for some $t > 5$ we have $\gamma_t = \gamma_{t-1} = ... = \gamma_{t-5}$ then stop. Else, $t := t + 1$ and go to Step 1.

The values for the parameters M, ρ and α reported here are in line with the ones suggested in the literature (Rubinstein and Kroese, 2004) for combinatorial problems such as this one. After the execution of the algorithm, the updated vector of position probabilities converges to sequence of '0's and '1's, which corresponds to the best segmentation under g.

The extension to bilingual level is done by incorporating the structural quality measure of Section 3. The setting is similar, i.e., samples are again bit-strings, but of length $\ell = n + m - 2$, where n and m are the number of words in the source and target sentence respectively. The first $n - 1$ bits correspond to the source sentence and the rest to the target sentence. The surface quality score of such a bit-string is given by the harmonic mean of its source and target surface quality scores.[1] The bit-string scoring function throughout Steps 1 – 3 is given by the harmonic mean of surface and structural quality scores. Finally, N-best lists are trivially generated, simply by collecting the top-N performing accumulated samples of a maximization process.

5 Experiments

Given a sentence pair with known and fixed word alignments, the result of the method described in Section 4 is an N-best list of bilingual segmentations of such a pair. The objective function provides a balance between compositional expressive power of segments in both languages and synchronization via word alignments. Thus, each (continuous) component of such a bilingual segmentation leads to the extraction of a high quality phrase pair.

As was mentioned in Section 3, each extracted phrase pair of standard phrase-based SMT is constructed from a component or from the union of components of an unsegmented word-aligned sentence pair. For each sentence pair, all possible (continuous) components and (continuous) unions of components give rise to the extracted (continuous) phrase pairs. In this section we investigate the impact to SMT models and translation quality, when extracting phrase pairs (from the N-best lists) that correspond to components *only*. A reduction in phrase-table size is guaranteed because we are essentially extracting only a subset of all possible continuous phrase pairs. The challenge is to verify whether this subset can provide a sufficient translation model.

[1] As it was mentioned in Section 2 the surface quality score in (6) is a real number. At each iteration of the algorithm the surface score of a segmentation can be converted into a number in $[0, 1]$ via Min-Max normalization. This holds for both source and target sides of a bit-string (independently).

	Cz–En	De–En
Europarl (v7)	642,505	1,889,791
News Commentary (v8)	139,679	177,079
Total	782,184	2,066,870

Table 1: Number of filtered parallel sentences for Czech–English and German–English.

Both the baseline and our system are standard phrase-based MT systems. Bidirectional word alignments are generated with GIZA++ (Och and Ney, 2003) and 'grow-diag-final-and'. These are used to construct a phrase-table with bidirectional phrase probabilities, lexical weights and a reordering model with monotone, swap and discontinuous orientations, conditioned on both the previous and the next phrase. 4-gram interpolated language models with Kneser-Ney smoothing are built with SRILM (Stolcke, 2002). A distortion limit of 6 and a phrase-penalty are also used. All model parameters are tuned with MERT (Och, 2003). Decoding during tuning and testing is done with Moses (Koehn et. al, 2007). Since our system only affects which phrases are extracted, lexical weights and reordering orientations are the same for both systems.

Datasets are from the WMT'13 translation task (Bojar et al., 2013): Translation and reordering models are trained on Czech–English and German–English corpora (Table 1). Language models and segment measures $gain$, as defined in (5), are trained on 35.3M Czech, 50.0M German and 94.5M English sentences from the provided monolingual data. Tuning is done on newstest2010 and performance is evaluated on newstest2008, newstest2009, newstest2011 and newstest2012 with BLEU (Papineni et al., 2001).

In our experiments the size of an N-best list varies according to the total number of words in the sentence pair, say w. For the purposes of phrase extraction in SMT we would ideally require all local maxima to be part of an N-best list. This would

Method	Czech→English				English→Czech				Czech–English
	'08	'09	'11	'12	'08	'09	'11	'12	PT size (retain%)
Baseline	19.6	20.6	22.6	20.6	14.8	15.6	16.6	14.9	44.6M (100%)
N-best	19.7	20.4	22.4	20.3	14.4	15.2	16.3	14.3	4.4M (9.8%)
N-best & unseg.	19.6	20.5	22.6	20.7	14.6	15.4	16.8	14.7	4.6M (10.4%)

Table 2: BLEU scores and phrase-table (PT) sizes for Czech–English. Phrase-table of 'Baseline' is constructed from all consistent phrase pairs. Phrase-table of 'N-best' is constructed from consistent phrase pairs that are components of the top-N bilingual word-aligned segmentations of each sentence pair. Similarly for 'N-best & unseg.', but consistent phrase pairs that are components of each (unsegmented) sentence pair are also included.

Method	German→English				English→German				German–English
	'08	'09	'11	'12	'08	'09	'11	'12	PT size (retain%)
Baseline	21.4	20.8	21.3	22.1	15.1	15.1	16.0	16.5	102.3M (100%)
N-best	21.3	20.6	21.3	21.8	15.0	15.0	15.6	16.0	9.4M (9.2%)
N-best & unseg.	21.5	20.8	21.5	22.0	15.4	15.2	15.7	16.2	9.9M (9.7%)

Table 3: Similar to Table 2, but for German–English.

guarantee the extraction of all high quality phrase pairs, with (empirically) desired variations, while keeping N small. Since the CE method performs global optimization, the resulting members of an N-best list are in the vicinity of the global maximum. Consequently, we cannot guarantee the inclusion of local maxima. We set $N = \lceil 30\%w \rceil$ so that at least some variation from the global maximum is included, but is not large enough to contaminate the lists with noisy bilingual segmentations. The resulting lists have 22 bilingual segmentations on average for both language pairs. Figure 4 shows typical German–English best performing bilingual segmentations.

BLEU scores are reported in Tables 2 and 3 for Czech–English and German–English respectively. Methods 'Baseline' and 'N-best' are the ones described above. Phrase-table sizes are reduced as expected and performance when translating to English is comparable. The significant drops in newstest2012 when translating from the morphologically poorer language (English) prompts us to include more 'basic' phrase pairs in the phrase-tables. This leads to augmenting each N-best list by its unsegmented sentence pair. Consequently, method 'N-best & unseg.' extracts the same phrase pairs as 'N-best', together with those from components of the unsegmented sentence pairs. As a result, translation quality is comparable to 'Baseline' across all language directions and small phrase-table sizes are retained.

6 Discussion and Future Work

This work can also be viewed as an attempt to understand bilinguality as a *generalization* of monolinguality. There is conceptual common ground on what $gain(x)$ for phrase x (Section 2) or component x (Section 3) computes. In both cases it measures how 'stable' a unit is. The stability of a phrase x is determined by how difficult it is to split x into multiple phrases. The partially ordered set framework of partition refinements is the natural setting for such computations. In order to determine the stability of a component we turn to empirical evidence from SMT: 'good' phrase pairs are extracted from components or unions of components of the graph that represents word-aligned sentence pairs. The stability of a component x is therefore determined by how difficult it is to break x into multiple components. It is thus interesting to investigate whether there exists a general approach that unifies partition refinements and network reliability for the purpose of identifying highly stable multilingual units.

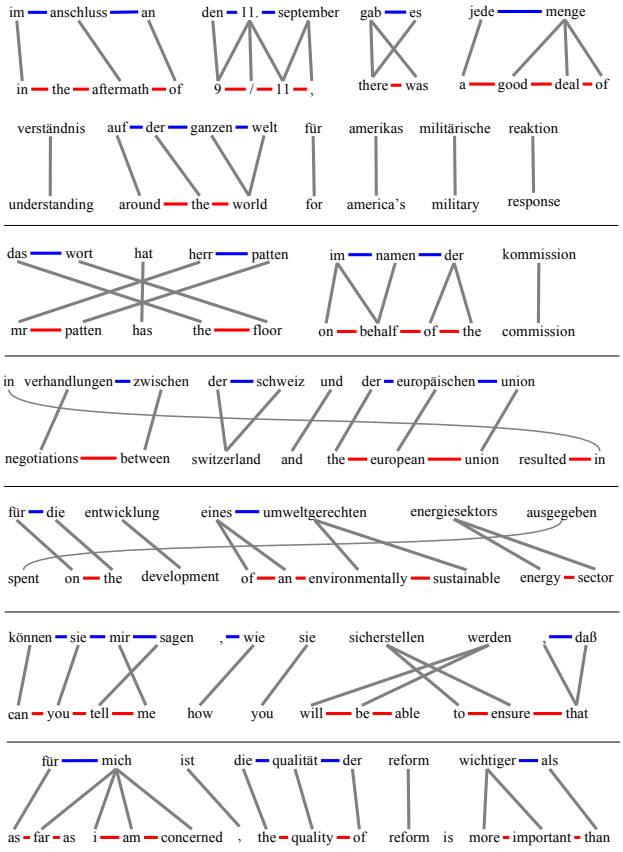

Figure 4: Typical fragments from best performing German–English segmentations.

The focus has been on bilingual segmentations, but as was mentioned in Section 2, it is possible to apply the CE method for generating monolingual segmentations. By using (6) as the objective function, we observed that the resulting segmentations yield promising applications in n-gram topic modeling, named entity recognition and Chinese segmentation. However, in the spirit of Ries et al. (1996), attempts to minimize perplexity instead of maximizing (6), resulted in larger segments and the segment quality definition of Section 1 was not met.

The sizes of the resulting phrase-tables together with the type of phrase pairs that are extracted lead to applications involving discontinuous phrase pairs. In (Galley and Manning, 2010) there was evidence that discontinuous phrase pairs that are extracted from discontinuous components of word-aligned sentence pairs can improve translation quality.[1] As the number of such components is much bigger than the continuous ones, (Gimpel and Smith, 2011) propose a Bayesian nonparametric model for finding the most probable discontinuous phrase pairs. This can also be done from the N-best lists that are generated in Section 4, and it would be interesting to see the effect of such phrase pairs in our existing models.

In a longer version of this work we intend to study the effect in translation quality when varying some of the parameters (size of N-best lists, sample sizes for training $gain$ in Section 2 and for the CE method), as well as when extracting source-driven bilingual segmentations as in (Sanchis-Trilles et al., 2011).

7 Conclusions

In this work, we have presented a solution to the problem of extracting bilingual segmentations in the presence of word alignments. Two measures that assess the quality of bilingual segmentations based on the expressive power of segments in both languages and their synchronization via word alignments have been introduced. We have established the link between the CE method and finding the best monolingual and bilingual segmentations. These measures formed the objective function of the CE method whose maximization resulted in an N-best list of bilingual segmentations for a given sentence pair. By extracting only phrase pairs that correspond to components from bilingual segmentations of those lists, we found that phrase table sizes can be reduced with insignificant loss in translation quality.

Acknowledgements

This research was funded in part by the European Commission through the CoSyne project FP7-ICT-4-248531 and the Netherlands Organisation for Scientific Research (NWO) under project nr. 639.022.213.

[1] By 'discontinuous component' we mean a component whose source or target words (vertices) form a discontinuous substring in the source or target sentence respectively.

References

Graeme Blackwood, Adria de Gispert, and William Byrne. 2008. Phrasal Segmentation Models for Statistical Machine Translation. In *COLING*.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *WMT*.

Charlie J. Coulbourn. 1987. The Combinatorics of Network Reliability. Oxford University Press.

Pieter-Tjerk De Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. 2005. A Tutorial on the Cross-Entropy Method. *Annals of Operations Research*, vol. 134, pages 19–67.

Michel Galley and Christopher D. Manning. 2010. Accurate Non-Hierarchical Phrase-Based Translation. In *NAACL*.

Kevin Gimpel and Noah A. Smith. 2011. Generative Models of Monolingual and Bilingual Gappy Patterns. In *WMT*.

Jin Guo. 1997. Critical Tokenization and its Properties. *Computational Linguistics*, vol. 23(4), pages 569–596.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrase-table. In *EMNLP-CoNLL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *ACL, demonstration session*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *HLT-NAACL*.

Spyros Martzoukos, Christophe Costa Florêncio, and Christof Monz. 2013. Investigating Connectivity and Consistency Criteria for Phrase Pair Extraction in Statistical Machine Translation. In *Meeting on Mathematics of Language*.

Bernard Monjardet. 1981. Metrics on partially ordered sets – a survey. *Discrete Mathematics*, vol. 35, pages 173–184.

Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, vol. 29 (1), pages 19–51.

Franz J. Och, Christoph Tillmann, and Hermann Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In *EMNLP-VLC*.

Chris Orum and Cliff A. Joslyn. 2009. Valuations and Metrics on Partially Ordered Sets. *Computing Research Repository - CORR*, vol. abs/0903.2.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2010. Integration of Multiple Bilingually-Learned Segmentation Schemes into Statistical Machine Translation. In *WMT and MetricsMATR*.

Klaus Ries, Finn Dag Bu, and Alex Waibel. 1996. Class phrase models for language modeling. In *ICSLP*.

Reuven Y. Rubinstein. 1997. Optimization of Computer Simulation Models with Rare Events. *European Journal of Operations Research*, vol. 99, pages 89–112.

Reuven Y. Rubinstein and Dirk P. Kroese. 2004. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning. Springer-Verlag, New York.

Germán Sanchis-Trilles, Daniel Ortiz-Martínez, Jesús González-Rubio, Jorge González, and Francisco Casacuberta. 2011. Bilingual segmentation for phrasetable pruning in Statistical Machine Translation. In *EAMT*.

Richard P. Stanley. 1997. Enumerative Combinatorics, Volume 1. Cambridge University Press.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *ICSLP*.

Leslie G. Valiant. 1979. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, vol. 8, pages 410–421.

Dominic J. A. Welsh. 1997. Approximate counting. *Surveys in Combinatorics*, London Math. Soc. Lecture Notes Ser., 241, pages 287–324.

Richard Zens, Daisy Stanton, and Peng Xu. 2012. A Systematic Comparison of Phrase Table Pruning Techniques. In *EMNLP*.