Collaborative Matching for Sentence Alignment*

Xiaojun Quan[1], Chunyu Kit[2], and Wuya Chen[1]

[1] School of Data and Computer Science, Sun Yat-sen University
quanxj3@mail.sysu.edu.cn,chenwy58@mail2.sysu.edu.cn
[2] Department of Linguistics and Translation, City University of Hong Kong
ctckit@cityu.edu.hk

Abstract. Existing sentence alignment methods are founded fundamentally on sentence length and lexical correspondences. Methods based on the former follow in general the length proportionality assumption that the lengths of sentences in one language tend to be proportional to that of their translations, and are known to bear poor adaptivity to new languages and corpora. In this paper, we attempt to interpret this assumption from a new perspective via the notion of collaborative matching, based on the observation that sentences can work collaboratively during alignment rather than separately as in previous studies. Our approach is tended to be independent on any specific language and corpus, so that it can be adaptively applied to a variety of texts without binding to any prior knowledge about the texts. We use one-to-one sentence alignment to illustrate this approach and implement two specific alignment methods, which are evaluated on six bilingual corpora of different languages and domains. Experimental results confirm the effectiveness of this collaborative matching approach.

Keywords: Sentence alignment · Machine translation.

1 Introduction

Sentence alignment has been extensively studied as a first step towards more ambitious natural language processing tasks such as statistical machine translation [2] and cross-language information retrieval [12]. The task is to identify translational correspondence between bilingual sentences in parallel text, also called bitext. The correspondence can then be taken as input to, for example, produce translational correspondence between bilingual words or phrases so as to build a machine translation model. Besides sentence length, lexical clues are also resorted to facilitate sentence alignment, yet they are not necessarily available from scenario to scenario. Thus, a main stream of research in text alignment remains not counting on lexical information but instead on the exploitation of length proportionality between mutual translations. The proportionality follows the observation that longer sentences in one language are likely to be translated into longer sentences in another language, and so are short ones. Statistical evidence also validates that the correlation of sentence lengths between two languages is relatively high [5].

* The paper was supported by the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (No.2017ZT07X355).

Table 1: Estimation of normal distribution parameters (mean and variance) across corpora of different languages.

		Eng-Chi$_1$	Eng-Chi$_2$	Eng-Fre	Eng-Ger	Eng-Por	Eng-Spa
Mean	char	0.520	0.386	1.156	1.188	1.108	1.097
	word	1.066	0.868	1.060	0.940	1.026	1.054
Var	char	0.045	0.029	0.300	0.180	0.040	0.041
	word	0.110	0.118	0.050	0.040	0.049	0.051

Sentence length based alignment was initially studied in cognate languages [5] [1], and then applied to non-cognate languages [16]. Although the length correlation for non-cognate languages is not as high as that between cognate languages, this length based approach works fairly well. However, a drawback of length-based alignment arises from its insufficient adaptivity to new languages and corpora, because the relied distribution parameters have to be estimated on a bilingual corpus of two specific languages and are not suitable for others, especially unpopular languages. Table 1 illustrates this with distributional parameters estimated on six bilingual corpora, including two English-Chinese corpora and four corpora of English and French, German, Portuguese, and Spanish, covering various domains such as legislation, news and proceedings. Sentence lengths are measured in number of characters and words. The result shows that these parameters vary significantly not only across different language pairs but also across different corpora of the same languages pair. In practice, they need to be specifically customized towards suitable corpora in order to enable the alignment to work effectively.

Inspired by the same proportionality assumption, this paper attempts to go beyond the conventional way and propose the notion of collaborative matching to model length proportionality during alignment. This novel notion takes into account the synergies between sentences, measuring how likely two sequences of sentences should be aligned, instead of treating them separately as in previous works. Based on this notion, two new approaches to one-to-one sentence alignment are developed, one as an approximate solution to exhaustive search and the other built upon differential collaborative similarity. Note that the collaborative matching idea also applies to many-to-many alignment after proper formulation, yet its implementation is much more complicated. That is why we choose to use one-to-one alignment to illustrate our idea in this paper. Besides, we also provide new insights into the measurement of sentence length and present a self-information based approach. The new alignment methods are evaluated on six bilingual corpora of different languages and domains. Experimental results validate their effectiveness regardless whichever sentence length measurements are used, indicating the stability and adaptivity of the new collaborative matching.

2 Problem Statement

This section introduces basic notations and notions of sentence alignment using sentence length, to lay a background for formulating the idea of collaborative matching.

2.1 Notation

A sentence alignment algorithm takes a corpus of bitexts as input, which is comprised of a set of source-language sentences, $\mathcal{S} = \{s_1, s_2, \ldots, s_M\}$, and another set of target-language sentences, $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$. Its task is to find the translational correspondences between these two sets of sentences. Let $\mathcal{L}_\mathcal{S} = \{l_{s_1}, l_{s_2}, \ldots, l_{s_M}\}$ and $\mathcal{L}_\mathcal{T} = \{l_{t_1}, l_{t_2}, \ldots, l_{t_N}\}$ represent the respective sets of sentence lengths. Let $\mathcal{A} = \{a_1, a_2, \ldots, a_T\}$ be an alignment of $\mathcal{S}$ and $\mathcal{T}$, where $a_i = (\mathcal{S}_i; \mathcal{T}_i)$ for $\mathcal{S}_i \subset \mathcal{S}$ and $\mathcal{T}_i \subset \mathcal{T}$ corresponds to a pairing and T is the number of pairings produced. In the previous work by [1], $\mathcal{S}_i$ and $\mathcal{T}_i$ are also called "bead". For many-to-many alignment type (e.g., 2-2), $\mathcal{S}_i$ and $\mathcal{T}_i$ are each composed of multiple sentences to be merged together as a "bead".

2.2 Sentence Length

As discussed above, there are two natural ways to measure sentence length, namely in number of characters vs. words. Basically, both ways focus on the surface length of sentence and drop other essential information such as word identities. Given that the basic objective of translation is to render the meaning of a text of one language in another language, the equivalence of meaning is the ultimate means for text alignment. Length-based approaches resort to sentence length as a source of "information" about the meaning, assuming that it can be retained to a certain degree after translation.

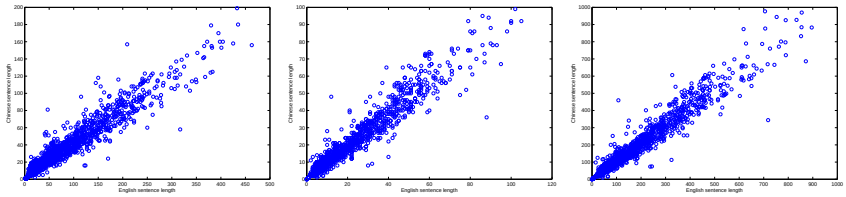
However, we argue that such "information" can be measured in a different way so as to make the best use of implicit lexical clues indicated by the distribution of words. Our assumption is that if two sentences are translation of one another, they should have a similar distribution of words in their respective texts. Accordingly, this distribution can be used to quantify the amount of "meaning" of a sentence. Specifically, suppose given a word w_i with probability of $Pr(w_i)$, we measure its length in terms of the amount of its self-information $-\log Pr(w_i)$ [4]. Analogically, the length of a sentence is the sum of accumulated self-information over all its words. The rationale of this new length measure can be illustrated by a statistical analysis on an English-Chinese corpus as in Figure 1, showing that this information based measure gives a higher correlation of sentence length than the other two.

2.3 Sentence Alignment

Most existing approaches to sentence alignment follow the monotonicity assumption that coupled sentences in bitexts appear basically in a similar sequential order in two languages, and employ dynamic programming for global optimization to produce the final output of alignment. Sentence alignment can be formulated as the following optimization

$$\mathcal{A}^* = \arg\max_{\mathcal{A}} \sum_{a_i \in \mathcal{A}} \log P(a_i), \tag{1}$$

where $P(a_i)$ is a probabilistic score measuring the likelihood of aligning sentences $\mathcal{S}_i$ and $\mathcal{T}_i$ in a_i in terms of their lengths. In [5] and [1], $P(a_i)$ is calculated by means of the sentence length ratio in two languages that is assumed to follow a normal distribution, with mean μ and variance σ^2 to be estimated on real data.



(a) Char (correlation = 0.9519) (b) Word (correlation = 0.9670) (c) Info (correlation = 0.9719)

Fig. 1: Correlation analysis of sentence length on an English-Chinese corpus. The length is respectively measured in number of characters and words and in amount of information.

3 Collaborative Matching

This section formulates our collaborative matching and illustrates its application to sentence alignment. A simple example is first given to illustrate the basic concept.

3.1 An Illustrative Example

Consider a scenario where we need to decide the translations of given sentences based solely on their lengths. Initially, suppose we are given a source-language sentence of length 4, as well as two candidates of translation with length 5 and 6, respectively. Without any additional knowledge, it is impossible to decide between the two candidates which is the potential translation. If one more sentence is added to each side of the three sets, say with respective length 5, 6, and 4, the answer becomes a bit clearer according to the length proportionality assumption. The source-language sentences (with length 4 and 5) are more likely to be translated into the target-language sentences with length 5 and 6 than that with 6 and 4. If another three sentences are given, say with lengths 6, 7 and 2, there should be little doubt about the correct alignment. Highest possibility is that source-language sentences with lengths $\{4, 5, 6\}$ are aligned with sentences with lengths $\{5, 6, 7\}$, rather than that with $\{6, 4, 2\}$.

From this example, we can see that sentences from the same language can work collaboratively during alignment rather than separately as in previous works. The collaboration is in fact a collective reflection of the length proportionality and can be intuitively demonstrated in the Euclidean space, in which the lengths of sentences from one language are assembled into a vector. The vector of true translations, that collectively satisfy the length proportionality better, have smaller angle (i.e., higher similarity) with the vector of source sentences. This gives a strong motivation for collaborative matching to measure how translations and their source texts fit each other.

3.2 Collaborative Similarity

The above example shows a new perspective on the use of length proportionality for sentence alignment. To this end, we propose the notion of collaborative similarity to measure the likelihood of aligning two equal-sized (this restriction will be released

later) sequences of sentences in terms of their length vectors. In particular, we propose to use a trigonometric function to estimate the similarity score. Among many possibilities, we use cosine function to produce a score monotonic with the likelihood. Specifically, the collaborative similarity of two sequences of sentences, $\hat{\mathcal{S}}$ and $\hat{\mathcal{T}}$, is defined as

$$C(\hat{\mathcal{S}}, \hat{\mathcal{T}}) = \frac{\sum_i l_{\hat{\mathcal{S}}_i} l_{\hat{\mathcal{T}}_i}}{\sqrt{\sum_i l^2_{\hat{\mathcal{S}}_i}} \sqrt{\sum_i l^2_{\hat{\mathcal{T}}_i}}}. \tag{2}$$

Then sentence alignment becomes a task of finding a set of aligned sentences with the maximum collaborative similarity from given bitext, as

$$(\hat{\mathcal{S}}, \hat{\mathcal{T}})^* = \arg\max_{(\hat{\mathcal{S}}, \hat{\mathcal{T}})} C(\hat{\mathcal{S}}, \hat{\mathcal{T}}). \tag{3}$$

Once $\hat{\mathcal{S}}$ and $\hat{\mathcal{T}}$ are aligned one by one sequentially to give the final 1-1 alignment, the remaining sentences will be treated as 1-0/0-1 type of alignment.

The differences of this approach from the previous alignment approaches are as follows. Given the prior knowledge (e.g., normal distribution) about the length proportionality, the previous approaches search for an alignment with the highest possibility. They can be regarded as a "supervised" alignment in the sense that their alignment results are produced subjecting to some prior knowledge. Our approach, however, is unable to align two single sentences straightforwardly but needs to rely on the synergies between sentences, which appears to be "unsupervised" in sense of using little prior knowledge.

3.3 Two Alignment Approaches

In our framework, sentence alignment becomes a task through exhaustive search of all possible alignments for one with the highest collaborative similarity, yet this is impractical due to the heavy computational cost. This section formulates two approximate approaches to resolving this issue.

Near-Optimal Alignment The first approach employs the divide-and-conquer strategy to find a near optimal solution through the following inference from Equation 3.

$$\begin{aligned}
(\hat{\mathcal{S}}, \hat{\mathcal{T}})^* &= \arg\max_{(\hat{\mathcal{S}}, \hat{\mathcal{T}})} \frac{\sum_{i=1}^{\hat{T}} l_{\hat{\mathcal{S}}_i} l_{\hat{\mathcal{T}}_i}}{\sqrt{\sum_{i=1}^{\hat{T}} l^2_{\hat{\mathcal{S}}_i}} \sqrt{\sum_{i=1}^{\hat{T}} l^2_{\hat{\mathcal{T}}_i}}} \\
&= \arg\max_{(\hat{\mathcal{S}}, \hat{\mathcal{T}})} \left(\frac{\sum_{i=1}^{\hat{T}-1} l_{\hat{\mathcal{S}}_i} l_{\hat{\mathcal{T}}_i}}{\sqrt{\sum_{i=1}^{\hat{T}} l^2_{\hat{\mathcal{S}}_i}} \sqrt{\sum_{i=1}^{\hat{T}} l^2_{\hat{\mathcal{T}}_i}}} + \frac{l_{\hat{\mathcal{S}}_{\hat{T}}} l_{\hat{\mathcal{T}}_{\hat{T}}}}{\sqrt{\sum_{i=1}^{\hat{T}} l^2_{\hat{\mathcal{S}}_i}} \sqrt{\sum_{i=1}^{\hat{T}} l^2_{\hat{\mathcal{T}}_i}}} \right) \\
&\approx \arg\max_{(\hat{\mathcal{S}}, \hat{\mathcal{T}})} \left(\frac{\sum_{i=1}^{\hat{T}-1} l_{\hat{\mathcal{S}}_i} l_{\hat{\mathcal{T}}_i}}{\sqrt{\sum_{i=1}^{\hat{T}-1} l^2_{\hat{\mathcal{S}}_i}} \sqrt{\sum_{i=1}^{\hat{T}-1} l^2_{\hat{\mathcal{T}}_i}}} + \frac{l_{\hat{\mathcal{S}}_{\hat{T}}} l_{\hat{\mathcal{T}}_{\hat{T}}}}{\sqrt{\sum_{i=1}^{\hat{T}} l^2_{\hat{\mathcal{S}}_i}} \sqrt{\sum_{i=1}^{\hat{T}} l^2_{\hat{\mathcal{T}}_i}}} \right)
\end{aligned} \tag{4}$$

That is, the final alignment can be approximately derived from the alignment of the preceding steps and the sentence pair under consideration in the current step. Correspondingly, the alignment can be performed via dynamic programming as follows. Let $W(i, j)$ be the maximum collaborative similarity that can be derived between sentences

$s_1, ..., s_i$ and their translations $t_1, ..., t_j$. Following the idea of near-optimal alignment, it is the max of the following three cases:

$$W(i,j) = \max \begin{cases} W(i-1,j) & + & 0 \\ W(i,j-1) & + & 0 \\ W(i-1,j-1) & + & \dfrac{l_{\hat{\mathcal{S}}_{\hat{T}}} l_{\hat{\mathcal{T}}_{\hat{T}}}}{\sqrt{\sum_{i=1}^{\hat{T}} l^2_{\hat{\mathcal{S}}_i}} \sqrt{\sum_{i=1}^{\hat{T}} l^2_{\hat{\mathcal{T}}_i}}} \end{cases} \tag{5}$$

The first two cases have zero second terms because the second term of Equation 4 is 0 for a 1-0/0-1 alignment. The dynamic programming procedure is initialized with zero $W(\cdot, 0)$ and $W(0, \cdot)$ and then proceeds to fill every cell of matrix W. The final alignment is retrieved by tracing W backwards starting from W_{MN}.

Differential Collaborative Similarity The collaborative similarity of two sequences of adjacent sentences from two languages can actually be used to find a set of neatly aligned sentences. Yet the actual alignment is intertwined with 1-0/0-1 type of alignment, making it infeasible to perform the entire alignment by virtue of the collaborative similarity of adjacent sentences. Nevertheless, high-quality initial alignment can be used as certain "benchmark" to measure how likely two new bilingual sentences are to be aligned. To this end, we propose the differential collaborative similarity. Before that, we first define three items associated with the initial alignment $\hat{\mathcal{A}}$. Let $d_{\hat{\mathcal{A}}} = \sum_{\hat{a}_i \subset \hat{\mathcal{A}},\ \hat{\mathcal{S}}_i \in \hat{a}_i,\ \hat{\mathcal{T}}_i \in \hat{a}_i} l_{\hat{\mathcal{S}}_i} l_{\hat{\mathcal{T}}_i}$, $q_{\hat{\mathcal{A}}} = \sum_{\hat{a}_i \subset \hat{\mathcal{A}},\ \hat{\mathcal{S}}_i \in \hat{a}_i} l^2_{\hat{\mathcal{S}}_i}$ and $r_{\hat{\mathcal{A}}} = \sum_{\hat{a}_i \subset \hat{\mathcal{A}},\ \hat{\mathcal{T}}_i \in \hat{a}_i} l^2_{\hat{\mathcal{T}}_i}$. Then, the differential collaborative similarity of two bilingual sentences, s_i and t_j, is defined as

$$w_{ij} = \frac{d_{\hat{\mathcal{A}}} + l_{s_i} l_{t_j}}{\sqrt{q_{\hat{\mathcal{A}}} + l^2_{s_i}} \sqrt{r_{\hat{\mathcal{A}}} + l^2_{t_j}}} - \frac{d_{\hat{\mathcal{A}}}}{\sqrt{q_{\hat{\mathcal{A}}}} \sqrt{r_{\hat{\mathcal{A}}}}}. \tag{6}$$

The rationale behind this measure is that the derived initial alignment is comprised of paired sentences highly reflecting the length proportionality. It is not unreasonable to assume that a new sentences pair should increase or at least preserve the collaborative similarity of an existing initial alignment if they are true translation of each other. This will contribute a positive differential collaborative similarity score. Otherwise, the score is negative. With this new similarity measure, the alignment can be performed by setting each $W(i,j)$ as

$$W(i,j) = \max \begin{cases} W(i-1,j) & + & \gamma \\ W(i,j-1) & + & \gamma \\ W(i-1,j-1) + w_{ij} & & \end{cases} \tag{7}$$

where γ is a penalty parameter, necessary in general in various sequence alignment tasks and has to be determined empirically. Similar dynamic programming as the above one for near-optimal alignment can be adopted straightforwardly to produce the final alignment using this similarity measure.

4 Evaluation

This section reports the evaluation of our new alignment approaches on six bilingual corpora, with Gale and Church's length based alignment approach as baseline for com-

parison, which is the most classic length-based alignment algorithm and serves as the foundation of most more advanced approaches and tools.

4.1 Data sets

The following six bilingual corpora of different languages and domains are used for evaluation.

BLIS. Bilingual Laws Information System (BLIS)[3] is an electronic text database of Hong Kong legislation. BLIS provides English-Chinese bilingual texts of ordinances and subsidiary legislation and organizes the legal texts into a hierarchy of chapters, sections, parts, subsections, paragraphs and subparagraphs. This corpus has been used in a number of previous works on sentence alignment and other machine translation tasks [7, 14]. 175 are randomly selected from the original 31,401 bitexts for manual alignment by two experts. Then, sentences are identified based on punctuations, resulting in a set of 1619 1-1 and 71 1-0/0-1 sentence pairs.

LDC. This corpus is comprised of English-Chinese news stories collected via Sinorama Magazine, Taiwan, from 1976 to 2004.[4] Of the original corpus of 365,568 sentence pairs, 27 bitexts of 2041 1-1 and 564 1-0/0-1 pairings are randomly selected for use in our evaluation.

Europarl Parallel Corpus. The original corpus is extracted from the proceedings of the European Parliament with versions in 21 European languages [8]. It is aligned at sentence level between English and other 20 languages. For this evaluation, 4 of the 20 aligned corpora of French, German, Portuguese, and Spanish are chosen. They are the cognate languages with the most widely used bitext data in the field. Each raw corpus includes more than 2 millions sentence pairs. Among them, 5000 pairs are randomly selected from each, with 1000 sentences from two sides are randomly discarded so as to form 1-0/0-1 type of alignment. Noted that there may be also pairs of 1-1 aligned sentences discarded during this process.

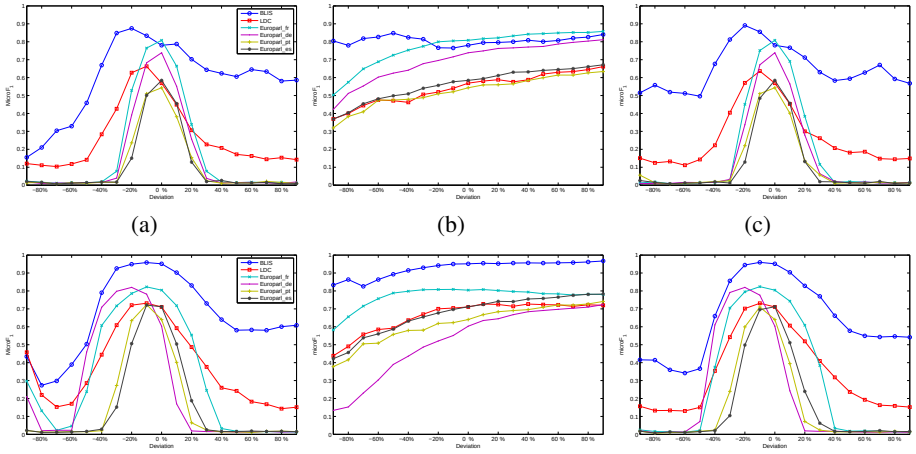
In our experiment, sentence length will be firstly measured in number of characters and number of word tokens. For the former, each cognate language character or punctuation is counted as 1, while each Chinese character is 2. For the latter, the length is measured in number of word tokens as segmented by spaces and punctuations. For Chinese, since there are no marked word boundaries, word segmentation needs to be performed. To measure sentence length in self-information of words, it is intuitively more desirable to stem them into their root forms, which, however, appears to be impractical as multiple languages are involved here while most existing stemming works focus only on the English language. For this reason, there will be no stemming performed for all the corpora.

4.2 Adaptivity of the Baseline

This part examines the adaptivity of Gale and Church's algorithm to the situation of inaccurate estimation of parameters. For this purpose, we first derive normal distribu-

[3] http://www.legislation.gov.hk

[4] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2005T10



(a) (b) (c)

(d) (e) (f)

Fig. 2: Adaptivity of Gale and Church's algorithm with different means (a, d), variances (b, e), and both (c, f). Sentence length in a-c and d-f is measured in number of characters and words, respectively.

tion parameters (mean and variance) from the six corpora according to the benchmark alignments and then alter them gradually by a rate of 10% each time. The alignment performance corresponding to such alteration of parameters is shown in Figure 2. It reveals the effect of different ways of changing the parameters: (a) and (d) resulting from changing mean only, (b) and (e) from changing variance only, and (c) and (f) from changing both. One observation we can make is that the algorithm is significantly more sensitive to the change of mean than that of variance. The possible reason is that the standardization of normal distribution makes the two parameters have different degrees of sensitivity. These figures also show that when the mean is misestimated by 20% or more, which could happen in practice, the alignment performance becomes disastrously poor. This study shows that the baseline tends to have weak adaptivity to the estimation of parameters.

4.3 Alignment by Collaborative Matching

This part reports the alignment result of near-optimal alignment (NOA) and differential similarity based alignment (DCSA), with sentence length measured in character, word and self-information. The size of the initial alignment for DCSA is set to $\min(M, N)/10$ of each bitext. Alignment performance is reported in precision (*P*), recall (*R*), defined as the proportions of correctly aligned pairings in produced pairings and gold standard, respectively, and F-measure (F_1), their harmonic mean, defined as $F_1 = 2PR/(P+R)$. In addition, micro-averaged performance in terms of precision, recall, and F-measure are also computed to measure the overall 1-1 and 1-0 performance. The final alignment results are shown in Table 2, 3 and 4, from which several findings can be derived. Firstly, Gale and Church's algorithm performs better with sentence

length in number of words than in number of characters, and suffers seriously from sentence length in amount of self-information. Secondly, although the non-parametric approach, NOA, achieves competitive performance against the baseline when sentence length is measured in number of characters, it underperforms when the length is measured in terms of words. But its performance is not prone to being affected by how sentence length is measured. Neither is the performance of DCSA, a stable and consistently outstanding performance across all the datasets.

Table 2: Alignment performance with sentence length in number of characters.

Dataset	Type	Gale&Church			NOA			DCSA		
		P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
BLIS	1-1	0.827	0.808	0.818	0.970	0.965	0.967	0.975	0.977	0.976
	1-0	0.180	0.380	0.244	0.565	0.732	0.638	0.686	0.676	0.681
	Micro	0.771	0.790	0.780	0.948	0.955	0.952	0.963	0.964	0.964
LDC	1-1	0.604	0.602	0.603	0.823	0.823	0.823	0.856	0.855	0.855
	1-0	0.441	0.452	0.447	0.678	0.681	0.680	0.676	0.681	0.678
	Micro	0.568	0.570	0.569	0.792	0.792	0.792	0.817	0.817	0.817
Ep_fr	1-1	0.847	0.877	0.862	0.689	0.698	0.693	0.875	0.901	0.888
	1-0	0.641	0.432	0.517	0.248	0.220	0.233	0.654	0.481	0.554
	Micro	0.821	0.796	0.808	0.617	0.611	0.614	0.845	0.824	0.834
Ep_de	1-1	0.790	0.814	0.802	0.719	0.729	0.724	0.861	0.880	0.870
	1-0	0.471	0.338	0.394	0.230	0.202	0.215	0.596	0.473	0.527
	Micro	0.748	0.729	0.739	0.642	0.635	0.639	0.823	0.808	0.815
Ep_pt	1-1	0.594	0.619	0.606	0.707	0.723	0.715	0.863	0.888	0.875
	1-0	0.241	0.148	0.183	0.263	0.209	0.233	0.627	0.457	0.529
	Micro	0.553	0.534	0.543	0.642	0.630	0.636	0.831	0.811	0.821
Ep_es	1-1	0.635	0.657	0.646	0.697	0.708	0.703	0.864	0.885	0.874
	1-0	0.302	0.208	0.247	0.266	0.227	0.245	0.639	0.497	0.559
	Micro	0.593	0.576	0.585	0.630	0.622	0.626	0.832	0.815	0.823

5 Related Work

Sentence length based alignment follows essentially the length proportionality assumption, which can be straightforwardly observed from Indo-European and non-Indo-European languages [16]. Sentence length, together with other information such as lexical correspondence, has served as the foundation of many successful sentence aligners. For example, [11] developed an aligner with a three-pass procedure, which first aligns bitext using only length information of sentences, from which a set of finely aligned sentence pairs is yielded for training a translation model [2]. Then, it realigns the bitext using both sentence length and the word correspondences derived by the trained model. Hunalign [15] is another aligner developed via a hybrid of sentence length and lexical information. When lexical information is unavailable, it performs a similar initial alignment using sentence length and then automatically generates a lexicon. If a lexicon

Table 3: Alignment performance with sentence length in number of words.

Dataset	Type	Gale&Church			NOA			DCSA		
		P	R	F_1	P	R	F_1	P	R	F_1
BLIS	1-1	0.968	0.968	0.968	0.970	0.962	0.966	0.981	0.981	0.981
	1-0	0.568	0.592	0.579	0.520	0.746	0.613	0.684	0.732	0.707
	Micro	0.950	0.952	0.951	0.943	0.953	0.948	0.968	0.970	0.969
LDC	1-1	0.753	0.758	0.756	0.798	0.791	0.794	0.837	0.842	0.839
	1-0	0.558	0.528	0.543	0.636	0.672	0.653	0.691	0.661	0.676
	Micro	0.713	0.709	0.711	0.761	0.765	0.763	0.806	0.803	0.805
Ep_fr	1-1	0.791	0.819	0.805	0.699	0.711	0.705	0.858	0.885	0.871
	1-0	0.528	0.358	0.427	0.257	0.216	0.235	0.617	0.446	0.517
	Micro	0.757	0.735	0.746	0.630	0.621	0.625	0.826	0.805	0.815
Ep_de	1-1	0.804	0.827	0.815	0.710	0.718	0.714	0.852	0.876	0.864
	1-0	0.539	0.394	0.455	0.240	0.215	0.227	0.617	0.451	0.521
	Micro	0.768	0.750	0.759	0.635	0.629	0.632	0.820	0.801	0.810
Ep_pt	1-1	0.775	0.805	0.790	0.703	0.718	0.711	0.859	0.882	0.87
	1-0	0.462	0.300	0.364	0.264	0.213	0.235	0.623	0.471	0.536
	Micro	0.737	0.714	0.725	0.638	0.627	0.632	0.826	0.808	0.817
Ep_es	1-1	0.813	0.841	0.826	0.697	0.709	0.703	0.864	0.883	0.874
	1-0	0.546	0.374	0.444	0.262	0.220	0.239	0.602	0.479	0.533
	Micro	0.779	0.757	0.768	0.630	0.621	0.626	0.826	0.811	0.818

Table 4: Alignment performance with sentence length in amount of self-information.

Dataset	Type	Gale&Church			NOA			DCSA		
		P	R	F_1	P	R	F_1	P	R	F_1
BLIS	1-1	0.857	0.837	0.847	0.976	0.967	0.971	0.981	0.981	0.981
	1-0	0.162	0.338	0.219	0.549	0.789	0.647	0.684	0.732	0.707
	Micro	0.797	0.816	0.806	0.950	0.960	0.955	0.968	0.970	0.969
LDC	1-1	0.477	0.470	0.473	0.817	0.816	0.817	0.862	0.866	0.864
	1-0	0.388	0.426	0.406	0.668	0.677	0.673	0.727	0.699	0.712
	Micro	0.456	0.461	0.458	0.785	0.786	0.785	0.833	0.830	0.831
Ep_fr	1-1	0.544	0.572	0.558	0.684	0.696	0.690	0.875	0.898	0.886
	1-0	0.253	0.137	0.178	0.249	0.212	0.229	0.637	0.484	0.55
	Micro	0.514	0.493	0.503	0.616	0.608	0.612	0.841	0.823	0.832
Ep_de	1-1	0.562	0.586	0.574	0.722	0.733	0.727	0.854	0.879	0.866
	1-0	0.226	0.136	0.170	0.243	0.207	0.224	0.626	0.453	0.526
	Micro	0.525	0.506	0.515	0.648	0.640	0.644	0.824	0.803	0.813
Ep_pt	1-1	0.572	0.602	0.587	0.708	0.724	0.716	0.855	0.876	0.865
	1-0	0.217	0.113	0.149	0.266	0.211	0.236	0.611	0.473	0.533
	Micro	0.537	0.514	0.525	0.643	0.631	0.637	0.820	0.803	0.812
Ep_es	1-1	0.540	0.565	0.552	0.716	0.731	0.723	0.865	0.885	0.874
	1-0	0.254	0.146	0.186	0.271	0.218	0.242	0.603	0.475	0.531
	Micro	0.509	0.490	0.499	0.650	0.639	0.644	0.827	0.811	0.819

exists, it yields a rough translation for a source text and then compares it with its true translation to form a similarity matrix. The matrix is then taken as input to a dynamic programming algorithm to produce a final alignment.

Besides internal lexical correspondences derived during alignment, many other works resort to external lexicons. For example, [6] leveraged an external dictionary together with an internally-derived lexicon to build lexical correspondences. [3] introduced a hybrid system for sentence alignment by combining sentence length and an external lexicon, as well as sentence offset information. To take fuller advantage of lexical information, [10] assumed that different words should have differentiated importances in his aligner - Champollion.[9] proposed a revised version of Champollion, to improve its speed without performance loss, by first dividing input bitext into smaller by a length-based approach and aligned fragments and then applying Champollion to derive finer-grained alignment. Another assumption that most approaches to sentence alignment follow is monotonicity assumption, that coupled sentences in bitexts appear in a similar sequential order in two languages. Differently, [14, 13] studied the problem of non-monotonic sentence alignment.

6 Conclusion

Sentence length has been widely utilized as a fundamental clue in most works and tools for sentence alignment. While most previous efforts have focused on resorting to probability theory to leverage this information under the length proportionality assumption, an unavoidable drawback is poor adaptivity to new languages and corpora. To find a solution, we establish our methodology on the notion of collaborative matching, an idea to follow this assumption in a collective manner to treat sentences collaboratively rather than separately during alignment. It makes the alignment less dependent on prior knowledge or specific languages and corpora and thus tends to be more adaptive. Furthermore, we have also provided new insights into the measurement of sentence length, presenting a new one by virtue of word self-information. Based on the idea of collaborative matching, we proposed two novel alignment methods and evaluated them on six corpora of different languages and domains. Several findings can be obtained from our experimental results. First, among the two conventional length measures, the one in number of words appears to be more reliable than the one in number of characters. Second, the proposed length measure leads to a stable and competitive performance with our alignment methods but a rather poor performance with the baseline. Next, one of the proposed alignment approaches, DCSA, has shown the best performance in most cases. Finally and more importantly, while the performance of the baseline is likely to be greatly affected by different measurements of sentence length, languages, or corpora, ours are consistently reliable, confirming the adaptivity of our methods.

References

1. Brown, P.F., Lai, J.C., Mercer, R.L.: Aligning sentences in parallel corpora. In: Proceedings of the 29th annual meeting on Association for Computational Linguistics (ACL '91). pp. 169–176 (1991)
2. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. Computational Linguistics **19**(2), 263–311 (1993)
3. Collier, N., Ono, K., Hirakawa, H.: An experiment in hybrid dictionary and statistical sentence alignment. In: Proceedings of the 17th International Conference on Computational Linguistics - the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL '98). pp. 268–274 (1998)
4. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience (1991)
5. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. In: Proceedings of the 29th annual meeting on Association for Computational Linguistics (ACL '91). pp. 177–184 (1991)
6. Haruno, M., Yamazaki, T.: High-performance bilingual text alignment using statistical and dictionary information. In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics (ACL '96). pp. 131–138 (1996)
7. Kit, C., Webster, J.J., Sin, K.K., Pan, H., Li, H., Kui, K., Haihua, S., Li, P.H.: Clause alignment for hong kong legal texts: A lexical-based approach. International Journal of Corpus Linguistics **9**, 29–51 (2004)
8. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT Summit 2005. pp. 79–86 (2005)
9. Li, P., Sun, M., Xue, P.: Fast-champollion: a fast and robust sentence alignment algorithm. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10): Posters. pp. 710–718 (2010)
10. Ma, X.: Champollion: A robust parallel text sentence aligner. In: LREC 2006. pp. 489–492 (2006)
11. Moore, R.C.: Fast and accurate sentence alignment of bilingual corpora. In: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users (AMTA '02). pp. 135–144 (2002)
12. Nie, J.Y., Simard, M., Isabelle, P., Durand, R.: Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In: Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99). pp. 74–81 (1999)
13. Quan, X., Kit, C.: Towards non-monotonic sentence alignment. Information Sciences **323**, 34–47 (2015)
14. Quan, X., Kit, C., Song, Y.: Non-monotonic sentence alignment via semisupervised learning. In: Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013). pp. 622–630 (2013)
15. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V.: Parallel corpora for medium density languages. In: Recent Advances in Natural Language Processing (RANLP 2005). pp. 590–596 (2005)
16. Wu, D.: Aligning a parallel english-chinese corpus statistically with lexical criteria. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL '94). pp. 80–87 (1994)