A Statistical Model for Measuring Structural Similarity between Webpages

Zhenisbek Assylbekov, Assulan Nurkas, Inês Russinho Mouga
School of Science and Technology
Nazarbayev University
53 Kabanbay batyr ave., Astana, Kazakhstan
{zhassylbekov, anurkas, ines.russinho}@nu.edu.kz

Abstract

This paper presents a statistical model for measuring structural similarity between webpages from bilingual websites. Starting from basic assumptions we derive the model and propose an algorithm to estimate its parameters in unsupervised manner. Statistical approach appears to benefit the structural similarity measure: in the task of distinguishing parallel webpages from bilingual websites our language-independent model demonstrates an F-score of 0.94–0.99 which is comparable to the results of language-dependent methods involving content similarity measures.

1 Introduction

A parallel corpus is a collection of text with translations into another language. Such corpora plays an important role in machine translation and multilingual language retrieval. Unfortunately, they are not readily available in the necessary quantities: some of them are subject to subscription or license fee and thus are not freely available, while others are domain-specific. However, there is the World Wide Web, which can be considered as one of the largest sources of parallel corpora, since there are many websites which are available in two or more languages. Many approaches have been therefore proposed for trying to exploit the Web as a parallel corpus: STRAND (Resnik and Smith, 2003), PT-Miner (Chen and Nie, 2000), BITS (Ma and Liberman, 1999), WPDE (Zhang et al., 2006), Bitextor (Esplà-Gomis and Forcada, 2010), ILSP-FC (Papavassiliou et al., 2013), etc. For most of these mining systems, there is a typical strategy for mining parallel texts: (1) locate bilingual websites; (2) identify parallel web pages; (3) extract bitexts. For the step (2) three main strategies can be found in the literature – they exploit:

- similarities in URLs;
- structural similarity of HTML files;
- content-similarity of texts.

Measuring structural similarity of HTML files, which is the “heart of STRAND” architecture (Resnik and Smith, 2003), involves calculating some quantitative features of candidate webpages and then comparing them to manually chosen threshold values or embedding those features into machine learning algorithms. Such approaches do not take into account the intrinsic stochastic nature of the mentioned features, and they require supervised learning of the parameters for each given website/language. In this paper we develop a more refined language-independent technique for measuring structural similarity between HTML pages, which uses the same amount of information as previous approaches, but is more accurate in distinguishing parallelism of webpages and can be applied in unsupervised manner.

2 Related Work

Measuring structural similarity between HTML files was first introduced in (Resnik, 1998), where a linearized HTML structure of candidate pairs was used to confirm parallelism of texts. Shi et al. (2006) used a file length ratio, an HTML tag similarity and a sentence alignment score to verify translational equivalence of candidate pages. Zhang et al. (2006) used file length ratio, HTML structure and content translation to train k-nearest-neighbors classifier for parallel pairs verification. Esplà-Gomis and Forcada (2010) used text-language comparison, file size ratio, total text length difference for preliminary filtering and then HTML tag structure and text block length were used for deeper filtering. In (San Vicente and Manterola, 2012) the bitext detection module runs

Proceedings of Recent Advances in Natural Language Processing, pages 24–31,
Hissar, Bulgaria, Sep 7–9 2015.

three major filters: link follower filter, URL pattern search, and a combination of an HTML structure filter and a content filter. In (Papavassiliou et al., 2013) structural filtering is based on length ratios and edit distances between linearized versions of candidate pairs. Liu et al. (2014) proposed a link-based approach in conjuction with content-based similarity and page structural similarity to distinguish parallel web pages from bi-lingual web sites.

To explain the essence of our work let us assume that candidate pairs are linearized as in STRAND and linearized sequences are aligned using a standard dynamic programming technique (Hunt and MacIlroy, 1976). For example, consider two documents that begin as follows:

```
<HTML>
<TITLE>The Republic of
Kazakhstan</TITLE>
<BODY>
<H1>The Republic of
Kazakhstan</H1>
The Republic of Kazakhstan is a unitary state
with a presidential form of
government.
⋮
```

```
<HTML>
<TITLE>Қазақстан
Республикасы</TITLE>
<BODY>
Қазақстан Республикасы
– президенттік басқару
нысанындағы біртұтас
мемлекет.
⋮
```

The aligned linearized sequences would be as follows:

[START: HTML]	[START: HTML]
[START: TITLE]	[START: TITLE]
[Chunk: 23]	[Chunk: 21]
[END: TITLE]	[END: TITLE]
[START: BODY]	[START: BODY]
[START: H1]	
[Chunk: 23]	
[END: H1]	
[Chunk: 72]	[Chunk: 69]

Let W denote the alignment cost, i.e. the total number of alignment tokens that are in one linearized file but not the other, M denote the total number of alignment tokens in one linearized file and N denote the total number of alignment tokens in the other linearized file (in the example above, $W = 3$, $M = 9$, $N = 6$). In all of the above-mentioned works the behavior of $W/(M+N)$ (or of W itself) is a crucial factor in making decision on parallelism of candidate pairs. However, the intrinsic stochastic nature of these quantities was never adressed before. In this paper we develop a statistical model for W, M and N, whose parameters can be estimated in unsupervised manner, and we show how structural filtering benefits from such model.

3 Statistical Model

3.1 Assumptions

Let random variables (r.v.) W, M, and N have the same meaning as in Section 2. Suppose that we are observing a pair of webpages for which $M = m$ and $N = n$. Then W is equal to the number of alignment tokens out of total $(m + n)$ tokens that are missing in either of the linearized sequences, which means that the r.v. W can be modeled by the binomial distribution with parameters $(m+n)$ and q, i.e.

$$\Pr(W = w | M = m, N = n) = \\ = \binom{m+n}{w} q^w (1-q)^{m+n-w}. \quad (1)$$

It is important to notice here that the parameter $q = \Pr(\text{token is removed})$ should be different for parallel and non-parallel pairs, since we expect significantly higher proportion of misalignments in non-parallel case than in parallel case. Thus, observing a small value of $W/(M + N)$ is one of the indicators in favor of parallelism of two pages. Another indicator is the similarity of M and N, which can be formalized in the following way:

$$N \begin{cases} = kM + b + \epsilon & \text{for a parallel pair,} \\ \text{indep. of } M & \text{for a non-parallel pair,} \end{cases} \quad (2)$$

where k, b are constants and the r.v. ϵ represents an error term of linear regression model, and is assumed to be independent from M and N. Our investigation shows that a Gaussian mixture model (GMM) fits well the distribution of ϵ (See Appendix A). Therefore we assume that ϵ is distributed according to the pdf

$$f_\epsilon(x; \lambda, \mu_{1,2}, \sigma_{1,2}) \\ = \frac{1}{\sqrt{2\pi}} \left(\frac{\lambda}{\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \frac{1-\lambda}{\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \right). \quad (3)$$

The third indicator of parallelism that we are going to exploit is the similarity between text lengths: if L_1 and L_2 denote total lengths of text chunks in a candidate pair of webpages, then we assume that

$$L_2 \begin{cases} = aL_1 + c + z\sigma\sqrt{L_1} & \text{for a par. pair,} \\ \text{indep. of } L_1 & \text{for a non-par. pair,} \end{cases} \tag{4}$$

where a, c, σ are constants, z is a standard normal random variable and the variance of the difference $(L_2 - aL_1 - c)$ is modeled proportional to the length L_1 as in (Gale and Church, 1993). We notice here, that the assumptions (1) and (2) were made regardless of the text lengths L_1 and L_2: thus knowing the values of L_1 and L_2 does not affect the distribution of W (when M and N are given) or the joint distribution of (M, N).

Hereinafter we use the following notation: $\hat{p}_X(x)$ denotes an empirical pdf for a r.v. X, calculated from a set of observations $\{x_i\}$; the symbol "$\|$" is used to denote that "pages under consideration are parallel"; and the symbol "$\nparallel$" is used to denote that "pages under consideration are not parallel". When there is no possibility for confusion, we write Pr(x) for Pr$(X = x)$, and use similar shorthands throughout.

3.2 Derivation

Let us denote $\boldsymbol{x} = (w, m, n, l_1, l_2)$. Our ultimate goal is to be able to calculate Pr$(\| \; |\boldsymbol{x})$ and Pr$(\nparallel \; |\boldsymbol{x})$, and then to compare them in order to select the most probable case. These probabilities can be rewritten using Bayes' rule:

$$\begin{aligned} \Pr(\| \; |\boldsymbol{x}) &= \frac{\Pr(\boldsymbol{x}| \; \|) \Pr(\|)}{\Pr(\boldsymbol{x})} \\ \Pr(\nparallel \; |\boldsymbol{x}) &= \frac{\Pr(\boldsymbol{x}| \nparallel) \Pr(\nparallel)}{\Pr(\boldsymbol{x})} \end{aligned} \tag{5}$$

Since the denominators in (5) are same, it is sufficient to compare the numerators. Now, let us derive a model for the distribution of W, M, N, L_1 and L_2 in case of a parallel pair:

$$\begin{aligned} A_{\|} :=& \Pr(w, m, n, l_1, l_2| \; \|) = \\ =& \Pr(w, m, n|l_1, l_2, \|) \Pr(l_1, l_2| \; \|) = \\ =& \Pr(w|m, n, l_1, l_2, \|) \Pr(m, n|l_1, l_2, \|) \times \\ & \times \Pr(l_1, l_2| \; \|) = \\ =& \{\text{independence assumptions}\} = \\ =& \underbrace{\Pr(w|m, n, \|)}_{B_{\|}} \underbrace{\Pr(m, n| \; \|)}_{C_{\|}} \underbrace{\Pr(l_1, l_2| \; \|)}_{D_{\|}}. \end{aligned} \tag{6}$$

From (1) and the remark after it, we can say that

$$B_{\|} = \binom{m+n}{w} q_{\|}^{w} (1 - q_{\|})^{m+n-w}, \tag{7}$$

where $q_{\|} = \Pr(\text{token is removed}| \; \|)$. Also, from the assumption (2) we get

$$\begin{aligned} C_{\|} &= \Pr(M = m, kM + b + \epsilon = n) \\ &= \Pr(M = m) \cdot \Pr(kM + b + \epsilon = n|M = m) \\ &\approx \{\text{continuity correction for } \epsilon\} \\ &\approx \hat{p}_M(m) \Pr(\epsilon \in n - km - b \pm .5|M = m) \\ &= \{\text{independence of } M \text{ and } \epsilon\} \\ &= \hat{p}_M(m) \cdot \Pr(\epsilon \in n - km - b \pm .5) \\ &= \hat{p}_M(m) \cdot \int\limits_{n-km-b-.5}^{n-km-b+.5} f_\epsilon(x; \lambda, \mu_{1,2}, \sigma_{1,2}) dx, \end{aligned} \tag{8}$$

where $f_\epsilon(x; \lambda, \mu_{1,2}, \sigma_{1,2})$ is defined by (3). From the assumption (4) we obtain

$$\begin{aligned} D_{\|} =& \Pr\left(L_1 = l_1, aL_1 + c + z\sigma\sqrt{L_1} = l_2\right) \\ =& \Pr(L_1 = l_1) \\ & \times \Pr\left(aL_1 + c + z\sigma\sqrt{L_1} = l_2|L_1 = l_1\right) \\ \approx& \{\text{continuity correction for } z\} \\ \approx& \hat{p}_{L_1}(l_1) \cdot \Pr\left(z \in \frac{l_2 - al_1 - c \pm .5}{\sigma\sqrt{l_1}}\right) \\ =& \hat{p}_{L_1}(l_1) \cdot \frac{1}{\sqrt{2\pi l_1}\sigma} \int\limits_{l_2-al_1-c-.5}^{l_2-al_1-c+.5} e^{\frac{-x^2}{2l_1\sigma^2}} dx. \end{aligned} \tag{9}$$

Combining (6), (7), (8) and (9) we obtain

$$\begin{aligned} A_{\|} \approx& \binom{m+n}{w} q_{\|}^{w} (1 - q_{\|})^{m+n-w} \\ & \times \hat{p}_M(m) \cdot \int\limits_{n-km-b-.5}^{n-km-b+.5} f_\epsilon(x; \lambda, \mu_{1,2}, \sigma_{1,2}) dx \\ & \times \hat{p}_{L_1}(l_1) \cdot \frac{1}{\sqrt{2\pi l_1}\sigma} \int\limits_{l_2-al_1-c-.5}^{l_2-al_1-c+.5} e^{\frac{-x^2}{2l_1\sigma^2}} dx. \end{aligned} \tag{10}$$

Similarly, let us derive a model for the distribution of W, M, N, L_1 and L_2 in case of a non-parallel pair:

$$\begin{aligned}A_{\nparallel} :=& \Pr(w,m,n,l_1,l_2|\nparallel)\\ =& \Pr(w,m,n|l_1,l_2,\nparallel)\Pr(l_1,l_2|\nparallel) =\\ =& \Pr(w|m,n,l_1,l_2,\nparallel)\Pr(m,n|l_1,l_2,\nparallel)\times\\ & \times\Pr(l_1,l_2|\nparallel) =\\ =& \{\text{independence assumptions}\} =\\ =& \underbrace{\Pr(w|m,n\nparallel)}_{B_{\nparallel}}\underbrace{\Pr(m,n|\nparallel)}_{C_{\nparallel}}\underbrace{\Pr(l_1,l_2|\nparallel)}_{D_{\nparallel}}.\end{aligned} \tag{11}$$

As discussed earlier, under non-parallelism we should assume probability of an alignment token to be removed $q_{\nparallel}$ to be different from $q_{\parallel}$ and thus:

$$B_{\nparallel} = \binom{m+n}{w} q_{\nparallel}^{w}(1-q_{\nparallel})^{m+n-w}. \tag{12}$$

Due to independence assumption between M and N (2) under non-parallelism we have:

$$\begin{aligned}C_{\nparallel} &= \Pr(M=m|\nparallel)\cdot\Pr(N=n|\nparallel)\\ &\approx \{\text{marginal pdf's do not depend on } \nparallel\}\\ &\approx \hat{p}_M(m)\cdot\hat{p}_N(n).\end{aligned} \tag{13}$$

And, similarly, from (4) we have

$$\begin{aligned}D_{\nparallel} &= \Pr(L_1=l_1|\nparallel)\cdot\Pr(L_2=l_2|\nparallel)\\ &\approx \hat{p}_{L_1}(l_1)\cdot\hat{p}_{L_2}(l_2).\end{aligned} \tag{14}$$

Now, from (11), (12), (13) and (14) we obtain

$$\begin{aligned}A_{\nparallel} \approx& \binom{m+n}{w} q_{\nparallel}^{w}(1-q_{\nparallel})^{m+n-w}\\ &\times \hat{p}_M(m)\cdot\hat{p}_N(n)\cdot\hat{p}_{L_1}(l_1)\cdot\hat{p}_{L_2}(l_2).\end{aligned} \tag{15}$$

Our model $A_{\parallel}(w, m, n, l_1, l_2; q_{\parallel}, k, b, \lambda, \mu_{1,2}, \sigma_{1,2}, a, c, \sigma)$ has 11 parameters ($q_{\parallel}$, k, b, λ, $\mu_{1,2}$, $\sigma_{1,2}$, a, c, σ), it receives the values of w, m, n, l_1, l_2 as input, and outputs the probability to observe such values under *parallelism*. The model $A_{\nparallel}(w,m,n,l_1,l_2;q_{\nparallel})$ has one parameter ($q_{\nparallel}$), it also receives the values of w, m, n, l_1 and l_2 as input, and outputs the probability to observe such values under *non-parallelism*. For the sake of simplicity we will denote

$$\begin{aligned}\boldsymbol{\theta}_{\parallel} &= (q_{\parallel}, k, b, \lambda, \mu_{1,2}, \sigma_{1,2}, a, c, \sigma),\\ p_{\parallel} &= \Pr(\parallel).\end{aligned}$$

3.3 Parameters Estimation

In order to show how expectation maximization (EM) algorithm (Dempster et al., 1977) can be used to estimate the parameters of our models let us assume that the set of candidate pairs consists of s pairs. Let us introduce the variables (for $i = \overline{1,s}$)

$$\alpha_i = \begin{cases}1, & \text{if i}^{\text{th}} \text{ pair is parallel}\\ 0, & \text{otherwise.}\end{cases}$$

Then the likelihood function for our data is given by

$$\begin{aligned}L(q_{\parallel,\nparallel}, k, b, \lambda, \mu_{1,2}, \sigma_{1,2}, \sigma, p_{\parallel}) &=\\ = C\prod_{i=1}^{s}&[A_{\parallel}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\parallel})p_{\parallel}]^{\alpha_i}\times\\ &\times[A_{\nparallel}(\boldsymbol{x}_i;q_{\nparallel})(1-p_{\parallel})]^{1-\alpha_i},\end{aligned} \tag{16}$$

where $C = \prod_{i=1}^{s}[\Pr(\boldsymbol{x}_i)^{-1}]$ is a constant w.r.t. parameters $\boldsymbol{\theta}$, $q_{\nparallel}$, and $p_{\parallel}$. According to Lemma B.1, the likelihood (16) is maximized w.r.t $\{\alpha_i\}$ if

$$\alpha_i = \begin{cases}1, & \text{if } A_{\parallel}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\parallel})p_{\parallel} >\\ & \quad > A_{\nparallel}(\boldsymbol{x}_i;q_{\nparallel})(1-p_{\parallel}),\\ 0, & \text{otherwise.}\end{cases} \tag{17}$$

The formula (17) is basically the decision rule for our task of binary classification of candidate pairs into parallel or non-parallel ones (assuming that we know the parameters of $A_{\parallel}$ and $A_{\nparallel}$). Now the essence of the EM algorithm (Algorithm 1) can be described as follows.

We first initilize parameters on line 1 using the following reasoning: $q_{\parallel}$ should be less than $q_{\nparallel}$ due to the comment after (1); N should be approximately equal to M for parallel pairs, therefore we take $k = 1$ and $b = 0$ as initial guesses; since we know almost nothing about the components of the Gaussian mixture in (3), we set $\lambda = 0.5$ and $\mu_{1,2} = 0$, however we can expect that one of the components should be responsible for larger deviations from the mean (i.e. for heavy tails), and thus we set $\sigma_2 > \sigma_1$; we choose initial values for $a = 1$, $c = 0$ and $\sigma = \sqrt{6.8}$ based on the suggestion in (Gale and Church, 1993), and for $p_{\parallel} = 2/3$ based on the experiments in (Resnik and Smith, 2003).

After such initial guesses on parameters, we perform an E-step on lines 3–10, i.e. the models $A_{\parallel}$ and $A_{\nparallel}$ are applied to the data, and as a result we obtain two sets of indexes: I keeps the indexes

Algorithm 1 EM algorithm for $A_{\parallel}$ and $A_{\nparallel}$

```
Input: set of values $\{(w_i, m_i, n_i, l_{1,i}, l_{2,i})\}_{i=1}^{s}$
Output: indexes $I \subset \{1, \ldots, s\}$ of parallel pairs,
    indexes $J \subset \{1, \ldots, s\}$ of non-parallel pairs,
    estimates for $q_{\parallel}$, $q_{\nparallel}$, $k$, $b$, $\lambda$, $\mu_{1,2}$, $\sigma_{1,2}$, $a$, $c$, $\sigma$,
    $p_{\parallel}$
 1: Initialize $q_{\parallel} \leftarrow 0.2$, $q_{\nparallel} \leftarrow 0.5$, $k \leftarrow 1$, $b \leftarrow 0$,
    $\lambda \leftarrow 0.5, \mu_1 \leftarrow 0, \mu_2 \leftarrow 0, \sigma_1 \leftarrow 1, \sigma_2 \leftarrow 10,$
    $a \leftarrow 1, c \leftarrow 0, \sigma \leftarrow \sqrt{6.8}, p_{\parallel} = 2/3.$
 2: while not converged do
 3:     for $i \in \{1, \ldots, s\}$ do
 4:         if $\frac{A_{\parallel}(\boldsymbol{x}_i;\boldsymbol{\theta}_{\parallel})}{1-p_{\parallel}} > \frac{A_{\nparallel}(\boldsymbol{x}_i;q_{\nparallel})}{p_{\parallel}}$ then
 5:             $\alpha_i \leftarrow 1$
 6:         else
 7:             $\alpha_i \leftarrow 0$
 8:         end if
 9:     end for
10:     $I \leftarrow \{i|\alpha_i = 1\}$, $J \leftarrow \{j|\alpha_j = 0\}$
11:     $q_{\parallel} \leftarrow \frac{\sum_{i\in I} w_i}{\sum_{i\in I}(m_i+n_i)}$
12:     $q_{\nparallel} \leftarrow \frac{\sum_{j\in J} w_j}{\sum_{j\in J}(m_j+n_j)}$
13:     $(k, b) \leftarrow \operatorname*{arg\,min}_{(k,b)} \sum_{i\in I} \rho(n_i - km_i - b)$
14:     for $i \in I$ do
15:         $\epsilon_i = n_i - km_i - b$
16:     end for
17:     $(\lambda, \mu_{1,2}, \sigma_{1,2}) \leftarrow$
            $\leftarrow \operatorname*{arg\,max}_{(\lambda,\mu_{1,2},\sigma_{1,2})} \prod_{i\in I} f_\epsilon(\epsilon_i; \lambda, \mu_{1,2}, \sigma_{1,2})$
18:     $(a, c) \leftarrow \operatorname*{arg\,min}_{(a,c)} \sum_{i\in I} \rho(l_{2,i} - al_{1,i} - c)$
19:     for $i \in I$ do
20:         $\delta_i = l_{2,i} - al_{1,i} - c$
21:     end for
22:     $\sigma \leftarrow \operatorname*{arg\,min}_{\sigma} \sum_{i\in I} \rho(\delta_i^2 - \sigma l_{1,i})$
23:     $p_{\parallel} \leftarrow |I|/s$
24: end while
```

of parallel pairs, and J keeps the indexes of non-parallel pairs. Then the M-step is performed on lines 11–23, where we update the parameters as follows: MLE for $q_{\parallel}$ and $q_{\nparallel}$ are given by Lemma B.2; the method of iteratively reweighted least squares is used to estimate k and b on line 13 where ρ is an Huber function (Huber, 2011). The obtained values for (k, b) are then used to calculate residuals $\{\epsilon_i\}_{i\in I}$; then, the parameters of GMM, $\lambda, \mu_{1,2}, \sigma_{1,2}$, are updated based on MLE (an additional EM-procedure is usually needed for this task); σ is estimated using robust linear regression (Huber, 2011) as suggested in (Gale and Church, 1993); finally, $p_{\parallel}$ is estimated as the proportion of parallel pairs.

An R-script, which implements the Algorithm 1, is available at `https://svn.code.sf.net/p/apertium/svn/branches/zaan/`.

4 Experiments

We selected five different websites to test our model: official site of the President of the Republic of Kazakhstan (`http://akorda.kz`), official site of the Ministry of Foreign Affairs of the Republic of Kazakhstan (`http://mfa.kz`), electronic government of the Repuplic of Kazakhstan (`http://egov.kz`), official site of the Presidency of the Portuguese Republic (`http://presidencia.pt`), and official site of the Prime Minister of Canada (`http://pm.gc.ca`). We downloaded all candidate pairs with the help of *wget* tool, and then removed boilerplates, i.e. navigational elements, templates, and advertisements which are not related to the main content, using simple Python scripts[1]. The details on the number of mined pairs are given in Table 1. We applied Algorithm 1 to all five websites (values of w, m, n, l_1, and l_2 were obtained using a modified version[2] of an open-source implementation of STRAND algorithm[3]). Then for each website we extracted a representative sample of candidate pairs and manually checked them (sample sizes were calculated based on Cochran's formula (Cochran, 2007) for

Website	Lang's	# of	Sample
		pairs	size
`akorda.kz`	kk-en	4135	352
`mfa.kz`	kk-en	180	180
`egov.kz`	kk-en	1641	312
`presidencia.pt`	pt-en	960	275
`pm.gc.ca`	fr-en	1397	302

Table 1: Websites for experiments

[1] the scripts as well as archives of the mined webpages are available at `https://svn.code.sf.net/p/apertium/svn/branches/kaz-eng-corpora`

[2] `https://github.com/assulan/STRANDAligner`

[3] `https://github.com/jrs026/STRANDAligner`

all websites except `mfa.kz`, for which we checked all pairs due to small amount of them). The metrics used to evaluate our model have been precision ($prec$), recall (rec), and F-score (F). The results of the experiments are given in Table 2.

Website	$prec$	rec	F
`akorda.kz`	0.941	0.971	0.956
`mfa.kz`	0.944	1.000	0.971
`egov.kz`	0.915	0.969	0.941
`presidencia.pt`	0.991	0.950	0.970
`pm.gc.ca`	0.990	1.000	0.995

Table 2: Results of the experiments

5 Discussion and Future Work

The experiments have shown that statistical modeling of misalignments in linearized HTML files allows us to get better results in the task of measuring structural similarity between webpages from bilingual websites. The previous approaches for measuring structural similarity were based on finding threshold values for the number of misalignments (W) or the misalignments ratio ($\frac{W}{M+N}$), or using these characterisics as features in machine learning algorithms. Those approaches either led to high precision but low recall, or required supervised learning of underlying models, or both. Our approach has good recall and acceptable precision rates; it is language-independent and the parameters of our model are estimated in unsupervised manner through EM algorithm.

We have noticed that the suggested algorithm demonstrates higher precision for websites, which have good quality of translated texts in general (e.g. `presidencia.pt`), than for websites, which have worse quality of translation (e.g. `egov.kz`); but it keeps recall at good level in all cases. This means that the model tries not to throw away parallel pairs, but it sometimes fails to recognize non-parallelism for the websites with substantial amount of medium or low quality of translated texts.

We now address the typical errors made by the model as well as possible directions for the future work. Type II errors (false negatives) are mainly caused by the pairs which have the same (or almost the same) content in two languages but there is significant difference in HTML-formatting of two pages (e.g. when *<p>* and *</p>* tags are used in one version to surround paragraphs, while the other version uses a sequence of *<br/><br/>* tags to separate paragraphs). This problem could be handled by an appropriate pre-processing (normalizing) of the HTML files before applying the Algorithm 1. Type I errors (false positives) are primarily caused by the pairs which are consistent in HTML-formatting but have some differences in content (e.g. when one or few sentences/short paragraphs are missing in one version but are present in the other version). This problem could be tackled by better alignment of text-chunks and better exploitation of the similarity in text lengths if we want to stay in a language-independent framework, or by embedding content-similarity measures, if we decide to switch to language-dependent techniques. In the latter case we could also use morphological segmentation as in (Assylbekov and Nurkas, 2014) for preprocessing texts in morphologically rich languages (like Kazakh), in order to improve the existing methods of measuring content-similarity.

Acknowledgments

We would like to thank Dr. Igor Melnykov for the useful discussions of the EM algorithm.

A Goodness-of-fit Tests for ϵ

Let r.v.'s W, M, and N be defined as in Section 2, and let w, m, and n denote values of these r.v.'s. We downloaded candidate pairs from the official website of the President of the Republic of Kazakhstan located at `http://akorda.kz` and then from each webpage we removed the boilerplate, i.e. navigational elements, templates, and advertisements which are not related to the main content[4]. For each candidate pair we obtained values of w, m, and n using a modified version[5] of an open-source implementation of STRAND algorithm[6]. The following heuristic rule was used to keep seemingly parallel pairs:

$$\{\text{pages are parallel}\} \approx \left\{\frac{W}{M+N} \in (0, 0.2]\right\} \cap \\ \cap \{M \in [19, 200]\} \cap \{N \in [19, 200]\}. \quad (18)$$

A threshold value of 0.2 for $W/(M+N)$ is recommended by the authors of STRAND. Boundaries for M and N are selected based on 1st and

[4] the scripts as well as the candidate pairs are available at `https://svn.code.sf.net/p/apertium/svn/branches/kaz-eng-corpora/akorda/`

[5] `https://github.com/assulan/STRANDAligner`

[6] `https://github.com/jrs026/STRANDAligner`

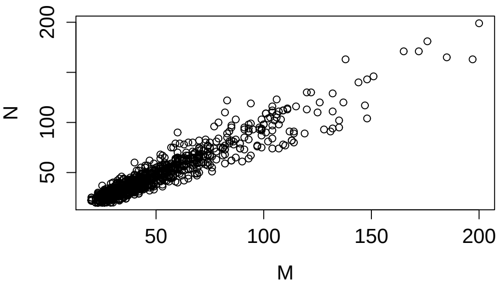

Figure 1: Scatter-plot of $\{(m_i, n_i)\}$ for seemingly parallel pairs.

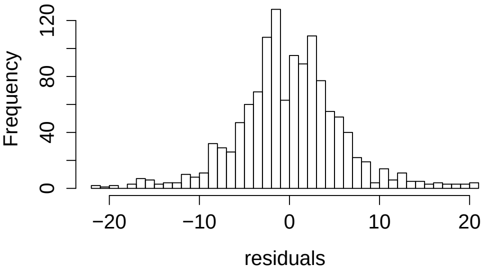

Figure 2: Distribution of the residuals $\{\epsilon_i\}$

99th percentiles and they are used to remove outliers. Application of the rule (18) resulted in 1271 seemingly parallel pairs. We stress here that the rule (18) *is not* used in our paper as the decision rule regarding parallelism of pages. Instead, it allows us to quickly identify pages which *seem* to be parallel and to look at the behavior of their M and N values. Figure 1 provides a scatter-plot of $\{(m_i, n_i)\}_{i=1}^{1271}$ for the filtered set of pages and it shows that the rule (18) supports our assumption on the linear relationship between M and N for parallel pages (2).

Next, we fit a linear regression model $N = kM + b + \epsilon$ to the data (m_i, n_i), and look at the residuals $\epsilon_i = n_i - km_i - b$ (Figure 2). Outliers among $\{\epsilon_i\}$ are dropped based on 1st and 99th percentiles, which resulted in 1245 observations (instead of 1271).

Further on we show that ϵ can be modeled using a Gaussian mixture model. A two-component mixture of Gaussian distributions has a pdf

$$f_{GMM}(x; \lambda, \mu_1, \sigma_1, \mu_2, \sigma_2) = \\ = \frac{1}{\sqrt{2\pi}} \left(\frac{\lambda}{\sigma_1} e^{-\frac{(x-\mu_1)^2}{\sigma_1^2}} + \frac{1-\lambda}{\sigma_2} e^{-\frac{(x-\mu_2)^2}{\sigma_2^2}} \right) \tag{19}$$

We first find MLE $\lambda^e, \mu_1^e, \sigma_1^e, \mu_2^e, \sigma_2^e$ for the parameters in (19) using EM-algorithm (Dempster et al., 1977), and then test a hypothesis

$$\begin{aligned} H_0 : \ f_\epsilon(x) =& f_{GMM}(x; \mu_1^e, \sigma_1^e, \mu_2^e, \sigma_2^e) \\ H_1 : \ f_\epsilon(x) \neq& f_{GMM}(x; \mu_1^e, \sigma_1^e, \mu_2^e, \sigma_2^e), \end{aligned}$$

using the chi-square goodness-of-fit test. The details are provided in the Table 3, from where we decide not to reject H_0, i.e. there is no evidence that the residuals are not distributed according to (19). In other words, *a Gaussian mixture model does a good job in modelling* $\{\epsilon_i\}$.

Interval	Obs. Freq.	Exp. Freq.
$(-\infty, -19]$	5	5.26
$(-19, -16]$	10	6.92
$(-16, -14]$	9	8.03
$(-14, -12]$	8	12.38
⋮	⋮	⋮
$(12, 14]$	16	12.97
$(14, 16]$	8	8.62
$(16, 19]$	10	7.55
$(19, +\infty)$	7	5.88
χ^2 = 19.023, df = 19, p-value = 0.4554		

Table 3: Fitting a Gaussian mixture model to $\{\epsilon_i\}$

B Auxiliary Lemmas

Lemma B.1. *Let* $f(\alpha_1, \ldots, \alpha_n) = \prod_{i=1}^n p_i^{\alpha_i} q_i^{1-\alpha_i}$, *where* $\alpha_i \in \{0, 1\}$ *and* $p_i, q_i \in [0, 1]$, $i = \overline{1, n}$. *Then* f *reaches its maximum at*

$$\alpha_i = \begin{cases} 1, & \text{if } p_i > q_i \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

Proof. The proof is left as an excercise. □

Lemma B.2. *Let* $X_1, X_2, \ldots, X_m$ *be independent binomial random variables with parameters* (n_1, q), (n_2, q), ..., (n_m, q) *correspondingly. Then the maximum likelihood estimator for* q *is*

$$\hat{q} = \frac{\sum_{i=1}^m X_i}{\sum_{i=1}^m n_i} \tag{21}$$

Proof. The proof is left as an excercise. □

References

Zhenisbek Assylbekov and Assulan Nurkas. Initial explorations in kazakh to english statistical machine translation. In *The First Italian Conference on Computational Linguistics CLiC-it 2014*, page 12, 2014.

Jiang Chen and Jian-Yun Nie. Automatic construction of parallel english-chinese corpus for cross-language information retrieval. In *Proceedings of the sixth conference on Applied natural language processing*, pages 21–28. Association for Computational Linguistics, 2000.

William G Cochran. *Sampling techniques*. John Wiley & Sons, 2007.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Miquel Esplà-Gomis and Mikel Forcada. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93:77–86, 2010.

William A Gale and Kenneth W Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1993.

Peter J Huber. *Robust statistics*. Springer, 2011.

James Wayne Hunt and MD MacIlroy. *An algorithm for differential file comparison*. Bell Laboratories, 1976.

Le Liu, Yu Hong, Jun Lu, Jun Lang, Heng Ji, and Jianmin Yao. An iterative link-based method for parallel web page mining. *Proceedings of EMNLP*, pages 1216–1224, 2014.

Xiaoyi Ma and Mark Liberman. Bits: A method for bilingual text search over the web. In *Machine Translation Summit VII*, pages 538–542, 1999.

Vassilis Papavassiliou, Prokopis Prokopidis, and Gregor Thurmair. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, 2013.

Philip Resnik. *Parallel strands: A preliminary investigation into mining the web for bilingual text*. Springer, 1998.

Philip Resnik and Noah A Smith. The web as a parallel corpus. *Computational Linguistics*, 29(3): 349–380, 2003.

Inaki San Vicente and Iker Manterola. Paco2: A fully automated tool for gathering parallel corpora from the web. In *LREC*, pages 1–6, 2012.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. A dom tree alignment model for mining parallel data from the web. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 489–496. Association for Computational Linguistics, 2006.

Ying Zhang, Ke Wu, Jianfeng Gao, and Phil Vines. Automatic acquisition of chinese–english parallel corpus from the web. In *Advances in Information Retrieval*, pages 420–431. Springer, 2006.