CDA: a Cost Efficient Content-based Multilingual Web Document Aligner

Thuy Vu
Amazon Alexa AI
Manhattan Beach, CA, USA
`thuyvu@amazon.com`

Alessandro Moschitti
Amazon Alexa AI
Manhattan Beach, CA, USA
`amosch@amazon.com`

Abstract

We introduce a Content-based Document Alignment approach (CDA), an efficient method to align multilingual web documents based on content in creating parallel training data for machine translation (MT) systems operating at the industrial level. CDA works in two steps: (i) projecting documents of a web domain to a shared multilingual space; then (ii) aligning them based on the similarity of their representations in such space. We leverage lexical translation models to build vector representations using TF×IDF. CDA achieves performance comparable with state-of-the-art systems in the WMT-16 Bilingual Document Alignment Shared Task benchmark while operating in multilingual space. Besides, we created two web-scale datasets to examine the robustness of CDA in an industrial setting involving up to 28 languages and millions of documents. The experiments show that CDA is robust, cost-effective, and is significantly superior in (i) processing large and noisy web data and (ii) scaling to new and low-resourced languages.

1 Introduction

Online machine translation (MT) services require industrial-scale training data, i.e., significantly large and high-quality parallel sentences, to build accurate models. Exploiting the web for multilingual content has become a usual strategy in collecting large-scale parallel sentences for MT (Uszkoreit et al., 2010; Smith et al., 2013; Buck and Koehn, 2016a). Structural Translation Recognition for Acquiring Natural Data (STRAND) (Resnik and Smith, 2003) is a standard pipeline to extract parallel data from the web, consisting in three steps: (i) bilingual document alignment for an input set of documents, (ii) sentence alignment for each aligned document pair, and (iii) sentence filtering for non-translation or boilerplate cleaning. The first step

of identifying bilingual documents is technically challenging and made more complicated by the presence of large and noisy documents from web data.

In the WMT-16 Bilingual Document Alignment Shared Task (WMT16-BDAST), two standard approaches for identifying parallel pages[1] were studied: 1) URL matching heuristic (Smith et al., 2013) as a baseline and 2) content similarity as a solution to maximize the performance in identifying parallel documents. The benchmark on English-French document alignment task shows that the best top-1 recall (R@1) for each approach are 59.8% and 89.1%, respectively, as evaluated on the test set (Buck and Koehn, 2016a,b). The results, albeit conducted within English-French setting, indicate that leveraging document content can lead to a significant increase, up to 30 percent points in recall, and contributes ~50% novel bilingual document pairs.

The URL matching heuristic approach, named URL, identifies parallel pages using language identifiers, typically from ISO 639, annotated in the addresses. Pages, or web-documents, in different languages from a domain are aligned if their URLs are matchable after their language-identifier being removed (Smith et al., 2013). The strategy can identify a candidate at a decent cost by comparing two URLs without significant preprocessing needed. For example, the following URLs are a match: `xyz.ca/index.htm` and `xyz.ca/fr/index.htm` after removing "`fr/`" from the second URL. On the contrary, cost is the major issue when comparing content as it often requires language-specific processing and sophisticated modelings for cross-language normalization and alignment. The problem becomes even more challenging when dealing with web data and for

---

[1]"page" and "document" interchangeably refer to the content of a web page.

low-resourced languages.

We optimize the cost for the latter approach to enable its application at scale. Specifically, we design CDA to project multilingual documents to a shared multilingual space for direct similarity measurement. Therefore, we can run the STRAND pipeline for multiple languages at once to keep the pre-processing cost *monotonic* with respect to the number of languages and documents. In particular, we design CDA with two key objectives: i) minimal data processing cost and ii) fast scaling to new languages. The latter is also crucial to the language expansion in online MT services. Our contribution is three-fold:

- We present an optimized, efficient, and scalable framework that can perform multilingual document alignment at state-of-the-art performance in one go.

- To facilitate the development and evaluation, we created two web-scale datasets, which are much larger and have many more languages than what is currently publicly available.

- We study the contribution of CDA in multiple applications involving parallel sentence extraction for MT and mutual complement between URL and CDA.

We tested CDA with multiple large-scale datasets of web documents. We also studied the applications of CDA within an industrial setting, including (i) extracting more and better parallel sentences from an extremely large-scale dataset, (ii) producing better MT models, and (iii) improving yield, measured by the amount of extracted parallel content, for both URL and CDA in the STRAND pipeline. The experimental results show that, despite its minimality, CDA (i) is on par with the top systems from WMT16-BDAST, which use expensive bilingual resources, and (ii) can double the amount of parallel data extracted by previous URL-based approaches. Most importantly, we show that CDA provides robust performance when dealing with millions of documents and processing up to 28 languages, including low-resourced languages.

In the remainder of this paper, we summarize the previous work regarding document alignment in Section 2. We then describe the proposed system and our experiments in sections 3 and 4, respectively. Finally, we derive the conclusions of the paper in Section 5.

2 Related Work

Aligning multilingual documents is the key required processing in most multilingual text processing pipelines, including cross-lingual information retrieval (Steinberger et al., 2002; Pouliquen et al., 2004; Vulic and Moens, 2015; Jiang et al., 2020) and parallel data extraction. In the context of creating parallel training data for MT, the problem has been studied in the literature for comparable corpora (Munteanu et al., 2004; Vu et al., 2009; Pal et al., 2014) and web-structured parallel data extraction (Resnik, 1999; Uszkoreit et al., 2010; Buck and Koehn, 2016a). We focus on the latter in this paper.

WMT-16 Bilingual Document Alignment Shared Task is a recent shared-task focusing on identifying bilingual documents from crawled websites (Buck and Koehn, 2016b). The top 3 systems are **YODA** (Dara and Lin, 2016), **NOVALINCS** (Gomes and Pereira Lopes, 2016), and **UEDIN1 COSINE** (Buck and Koehn, 2016b). The first two require costly features, such as (i) n-gram comparison after translating all non-English text into English (Dara and Lin, 2016), and (ii) phrase-table of statistical MT (SMT) as a dictionary (Gomes and Pereira Lopes, 2016). **UEDIN1 COSINE** (Buck and Koehn, 2016b), on the other hand, only uses $TF \times IDF$-weighted with Cosine similarity. Interestingly, this method performs surprisingly well even without French-to-English translations, dropping just 3.4% in recall, from 93.7% to 90.3%. Though the finding can be due to the English and French lexicons' overlap, it suggests that $TF \times IDF$ with proper normalization is useful to compare document representations from sub-domains. Our proposed method exploits this aspect.

Given the advent of deep neural network modeling, document embeddings are among the main interests in general NLP applications (Le and Mikolov, 2014; Cer et al., 2018; El-Kishky and Guzmán, 2020). This line of research, however, is *not* technically related to our problem setting. Specifically, the cost to run a neural inference over a web-scale setting is prohibitively high, e.g., processing a dataset of several billion pages from CommonCrawl[2] is not feasible. To have an idea, the Cloud Translation[3] cost to translate a webpage having 20,000 characters is $0.4 as of Jan 2021.

---

[2]`commoncrawl.org`
[3]`cloud.google.com/translate/pricing`

3 Proposed System

We describe our method to identify parallel pages from a web domain. Specifically, pages in different languages are projected to a shared space, where their similarity can be measured.

Problem Definition Let $D = \{d_1, d_2, \ldots, d_n\}$ be the n pages from a domain, each page d_i is described by its content c_i in language L_i. The problem is to identify all (d_i, d_j) of different languages, $L_i \neq L_j$, and c_i and c_j are translational equivalent.

Multilingual Space Let $\mathcal{L} = \{L_1, L_2, \ldots\}$ be the set of languages found in D and let D_i be the set of documents in L_i. Thus, $D = \bigcup_{L_i \in \mathcal{L}} D_i$, where each D_i is associated with a lexicon V_i. Without loss of generality, we project documents from two languages, L_α and L_β, into a common space as follows. We first define alignment between two lexicons V_α and V_β as:

$$\mathcal{A} = \big\{ (a, b) : a \in V_\alpha, b \in V_\beta, \\ P_{\alpha \to \beta}(a, b) + P_{\beta \to \alpha}(b, a) \\ \geq P_{\alpha \to \beta}(a, w) + P_{\beta \to \alpha}(w, a), \forall w \in V_\beta \big\},$$

where $P_{\alpha \to \beta}$ and $P_{\beta \to \alpha}$ are lexical translation models from L_α to L_β, and vice versa[4]. It should be noted that $\mathcal{A}$ defines a common space, $\mathbb{R}^{|\mathcal{A}|}$, where the dimensions are all word pairs. However, we can simplify the approach by mapping all languages in the space of a pivot language, i.e., α. Thus, we define $\Pi_\alpha : D_\beta \longrightarrow \mathbb{R}^{|V_\alpha|}$ that maps documents $d_\beta \in D_\beta$ into the same space of D_α, as:

$$\Pi_\alpha(d_\beta) = \Pi_\alpha \left(w_1, w_2, \ldots, w_{|d_\beta|} \right) \\ = \vec{x} = \left(x_1, x_2, \ldots, x_{|V_\alpha|} \right) \quad (1)$$

We define a lexical mapping for document d_β as $M(d_\beta) = d_{\beta \to \alpha} = \{w'_1, ..., w'_{|d_{\beta \to \alpha}|} | (w_i, w'_j) \in \mathcal{A}\}$, which maps d_β into language α. Similar, we denote the mapping for a document collection D_β as $D_{\beta \to \alpha} = \{M(d_\beta) : \forall d_\beta \in D_\beta\}$. Finally, we compute the TF$\times$IDF representation of d_β as follows, $\forall w_i \in V_\alpha$:

- TF(w_i) = number of occurrences of w_i in $d_{\beta \to \alpha}$; and

---

[4]It can be easily shown that the proposed aligned is symmetric, i.e., the other condition $P_{\alpha \to \beta}(a, b) + P_{\beta \to \alpha}(b, a) \geq P_{\alpha \to \beta}(w, b) + P_{\beta \to \alpha}(b, w), \forall w \in V_\alpha$ holds.

- IDF$(w_i) = \log \left(1 + \frac{|D_{\beta \to \alpha}|}{1 + \#(w_i, D_{\beta \to \alpha})} \right)$, where $\#(w, D)$ returns the number of documents in D containing w.

We compute x_i in Eq. 1 using TF$(w_i) \times$ IDF(w_i).

Aligning Multilingual Documents Two documents are considered a good pair if their representations are similar, according to a similarity threshold t. We compute the similarity between d_i and d_j as the dot-product between, $v_i \cdot v_j \in [0..1]$ (we normalized the vector representations with ℓ_2). In practice, we use English to build the multilingual space as it is the dominant language on the Internet and in most multilingual websites.

4 Experiments

We examine the efficacy of CDA in this section. First, we describe (i) the pipeline setup for the experiments and (ii) our effort in creating suitable benchmark data and selecting relevant resources in Section 4.1, and Section 4.2, respectively. We then address the following performance aspects of CDA:

1. The performance in multilingual document alignment.

2. The impact of CDA, compared to URL, in an end-to-end STRAND pipeline.

3. The by-product applications of CDA in identifying (i) novel language identifiers beyond ISO 639 for URL and (ii) web-domains containing multilingual data that are not detectable using language identifiers.

4. The cost required to enable the support to a new language.

4.1 Pipeline Setup

Figure 1 depicts the STRAND pipeline for our experiments.

- The **input** is constituted by web documents of multiple domains. The **output** is a set of parallel sentence pairs extracted from the pipeline. Each document has a web address and a raw HTML source.

- The **pre-processing** step groups input documents by domain to create data for document alignment step using URL and CDA. For CDA, additionally, it extracts the
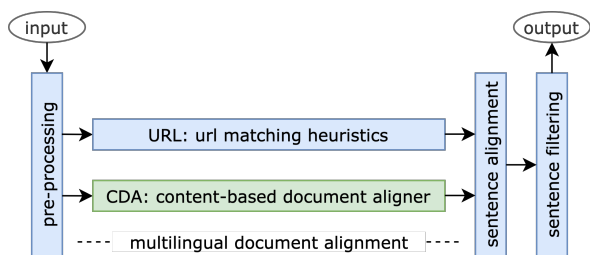
Figure 1: A STRAND pipeline that uses URL to align document and CDA as a replacement.

text content from HTML structure, using the following tags: `title, h1..h6, label, blockquote, dd, dt, p, pre, q, div`. This helps remove boiler-plate effectively from being considered in the calculation. We use Python's `langid` package to identify the language of a page.

- **Document alignment** is performed by either URL or CDA. For URL, we use a similar set of language identifiers from BDAST's base-line[5].

- For each aligned document pair, the **sentence alignment** step aligns text segments, called sentence pair candidates, of the aligned pages based on the DOM structure (Smith et al., 2013).

- Finally, the **sentence filtering** step removes low-quality pairs (Xu and Koehn, 2017; Sánchez-Cartagena et al., 2018) or duplica-tions. The filter we used in this experiment has approximately 90% F1 score for each lan-guage pair.

4.2 Dataset and Resource

We describe the datasets and resources used in the experiments.

4.2.1 Dataset

We collect and create the following datasets to study CDA performance in (i) matching parallel content, (ii) handling large datasets, and (iii) ex-tending its use to new languages.

WMT-16 Shared Task First, we use the bench-mark dataset provided for WMT-16 Shared Task on Bilingual Document Alignment. We evalu-ate and compare CDA with other English–French

document alignment methods on the BDAST's training set. The dataset consists of 348,858 and 225,043 English and French documents from 49 web-domains, respectively. Each document has a web address and a clean content. Besides, French documents are translated into English using a stan-dard SMT model. This translation is to study the potential upper-bound performance when having full translations. An alignment candidate has one document from each language, English or French, from the same domain. Thus, there are more than 4.2e9 possible alignments between the docu-ments. The golden data has 1,624 pairs provided by WMT16-BDAST. In this set, the number of la-beled alignments per domain ranges from 4 (e.g., `www.eohu.ca`) to 236 (e.g., `tsb.gc.ca`). The pairs generated by a system are first filtered by 1-1 rule: each document should participate in at most one alignment. A system is evaluated based on the recall achieved on these 1,624 pairs.

WMT-16 Deep Crawl The previous benchmark has two limitations. First, the size of the dataset is relatively small compared to a typical web-scale setting[6]. Second, the choice of English–French is not representative of the ultimate goal — find-ing more and better parallel data to enable MT in low-resourced languages. English–French has been the most studied pair in MT task. Besides, their lexicons are also highly overlapped (Lewis, 2009).

We address this problem, creating a larger dataset of more than 14MM pages using the same set of 49 domains. Specifically, we used these domains and URLs as seeds and recursively down-loaded all reachable pages from those seeds. We did not download pages that link to external do-mains. This exercise resulted in a dataset consisting of 8.7MM and 5.5MM pages for English and 28 other languages. These languages include: Arabic, Bulgarian, Chinese Simplified, Chinese Traditional, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Hebrew, Hungarian, In-donesian, Italian, Japanese, Korean, Malay, Norwe-gian, Polish, Portuguese, Russian, Slovak, Spanish, Swedish, Thai, and Turkish.

CommonCrawl Sextet Previous datasets share the same domains that are heavily biased toward French content (see Table 3). We leverage a monthly crawl from CommonCrawl, specifically

---

[5]https://github.com/christianbuck/wmt16-document-alignment-task/blob/master/languagestripper.py

[6]A typical multilingual domain can have thousands to millions of pages; e.g. `nato.int` and `microsoft.com` have 3e5 and 38e6 pages, respectively, indexed by Google[7].

Dataset	#Dom.	#EN-docs	#XX-docs	#$\mathcal{L}$
WMT-16 Shared Task	49	348,858	225,043	En+Fr
WMT-16 Deep Crawl	49	8.7MM	5.5MM	En+28
CommonCrawl Sextet	662	17.6MM	4.1MM	En+6

Table 1: Summary of Benchmark Datasets for Document Alignment Task. Each dataset is described by the number of web-domains (#Dom.), English documents (#EN-docs), non-English documents (#XX-docs), and languages (#$\mathcal{L}$).

CC-MAIN-2017-17[8], to create a better distribution dataset to validate CDA. We select pages from the crawl having pages in Chinese, Czech, Italian, Japanese, Russian, and Turkish. We only keep the following top-level domains: .cn, .tw, .cz, .it, .jp, .ru, .tr, .edu, .gov, and .org. The process results in a dataset of 600+ domains with 17.6MM and 4.1MM pages in English and six selected languages. Table 1 summarizes the datasets considered in our experiments.

4.2.2 Resource

Lexical translation dictionaries are the significant resource required in our proposed method to support a new language pair. Our experiments used the lexical translation dictionaries created by methods introduced for traditional SMT (Brown et al., 1993) and neural-based MT (Lample et al., 2018).

In particular, we use IBM-1 models for popular languages that have sufficient parallel data from general domains (Koehn, 2005). We use the GIZA++ toolkit (Och and Ney, 2003) to create IBM-1 models. Collecting such parallel data for the low-resourced languages is generally challenging. We instead leverage the advances in multilingual embeddings from the MUSE project[9] (Lample et al., 2018). We create translation probability between two words by their normalized embedding similarity score.

4.3 Bilingual Document Alignment Results

We evaluate the performance of CDA under the WMT-16 Shared Task benchmark. We conduct experiments on both settings, using the original text and using full translations. The latter setting allows us to understand the possible benefit of the expensive step, full document translation. Besides, the construction of V_{L_i} is crucial to the distinction of

---

---

baseline: URL		67.9
content-based systems	align w. org. text	align w. trans. text
YODA (Dara and Lin, 2016)	n/a	**93.71**
NOVALINCS (Gomes and Pereira Lopes, 2016)	90.50	n/a
UEDIN1 COSINE (Buck and Koehn, 2016b)	90.25[†]	93.65[†]
CDA: |V|=2,000	87.17	90.95
CDA: |V|=10,000	**90.40**	91.02
CDA: |V|=20,000	90.33	89.90

[†] *average performance of valid and test splits*

Table 2: Comparison of the baseline URL, top-3 performing systems from WMT16-BDAST and CDA under different settings of vocabulary size, using either original text (original documents in English and French) or translated text (documents in English and the English translations of documents in French)

the representations. Therefore, we examine the impact of the vocabulary size of V_{L_i} to the alignment result. Specifically, we construct V_{L_i} by selecting the top frequent tokens after removing stop-words and the first k frequent tokens. We empirically set k to be 100. We experiment with three different sizes for V_{L_i}: 2,000, 10,000, and 20,000. Finally, we compare the results of CDA with the baseline URL and the top-3 systems of the WMT-16 Bilingual Document Alignment Shared Task. The evaluation metric is the percentage of the 1,624 golden pairs found in the top-1 alignment for each English document. Table 2 shows the result.

For alignments using original text, the results indicate that CDA achieves similar performance with the state-of-the-art methods from BDAST. The result shows the efficacy of the proposed alignment method. The results also show that the vocabulary size, or the vector representations' size, impacts the performance. $|V| = 10,000$ yields the best result among the three settings. Second, even though using full translation is better, the performance gains are negligible with respect to the processing cost required for building the MT models and translating all the data into an anchor language. Since CDA does not exploit bi-gram features, its performance is relatively lower, up to 3%, compared to the state of the art. In short, the result suggests an optimal configuration for CDA with a vocabulary size of 10,000.

4.4 Multilingual Document Alignment Results

It was showed in WMT16-BDAST that content-based methods could add 60% more English–French document pairs compared to URL. This section aims to verify this in a multilingual setting, mainly when operating with a significantly higher number of languages and domains using WMT-16 Deep Crawl and CommonCrawl Sextet, respectively.

On WMT-16 Deep Crawl Table 3 shows the number of parallel documents and sentences extracted by an end-to-end STRAND pipeline, after filtering and duplication removal, described in Section 4.1. The result shows that CDA contributes an extra of 53%, 75%, and 195% in *clean* parallel sentences compared to URL for French alone, and when French is and is not considered, in this more extensive and more realistic setting, respectively. It also suggests that CDA is effective and can significantly increase the number of parallel sentences extracted for low-resourced languages. Finally, the result indicates that our proposed method is robust in a multilingual setting.

On CommonCrawl Sextet Table 4 shows the result in English parallel tokens extracted from the pipeline using URL and CDA in the document alignment step. The result shows similar gains as in the previous experiment, except for Czech — increasing 7× more parallel tokens. Our post-hoc analysis discovers non-standard language identifiers missing for URL processing, e.g., `ces` or `cesky`.

To confirm the study, we randomly selected 1,320 English–Turkish document pairs identified by CDA for human verification since we do not have annotated data. The outcome indicates that the accuracy of the document pairs is at 91.5%. Information on these datasets is described here: github.com/alexa/wqa_dataset.

4.5 Industrial Benchmarks

We conducted multiple internal experiments to examine the performance of CDA over URL under an industrial setting. Specifically, we focus on three application aspects of CDA: (i) robustness, (ii) identifying non-standard language identifiers for URL, (iii) identifying multilingual web-domains. Due to business security reasons, we do not name the specific languages considered in this study. We do

Language	# Document Pairs		# Sentence Pairs		
	URL	CDA	URL	CDA	CDA\URL
Arabic	3,266	2,896	36,262	39,590	24,065
Bulgarian	1,184	1,070	9,292	1,748	1,359
Chinese-S	2,805	2,160	30,519	27,666	13,289
Chinese-T	316	102	2,584	2,055	374
Croatian	704	3,119	889	56,300	55,854
Czech	29	241	77	7,264	7,248
Danish	137	2,932	693	39,488	38,996
Deutsch	5,525	8,863	83,663	170,932	113,851
Dutch	599	2,407	8,228	79,293	79,146
Farsi	1,316	1,404	14,697	13,875	6,122
Finnish	170	1,313	355	12,403	12,229
French	115,671	143,972	2,653K	3,568K	1,411K
Hebrew	209	140	7,742	5,295	83
Hungarian	1,253	1,382	10,494	6,158	4,448
Indonesian	368	551	625	1,204	900
Italian	6,644	7,310	57,977	94,098	55,802
Japanese	823	1,475	6,593	14,720	11,138
Korean	913	136	13,365	2,229	224
Malay	1,040	1,904	8,467	13,088	7,213
Norwegian	67	1,875	196	35,362	35,273
Polish	557	934	9,685	21,528	17,255
Portugese	1,545	6,200	12,294	104,561	96,850
Russian	3,984	2,475	36,565	55,010	40,722
Slovak	170	850	211	2,157	2,106
Spanish	8,334	21,765	114,874	317,430	252,523
Swedish	83	2,394	2,773	49,420	49,238
Thai	82	10	830	259	40
Turkish	1,057	2,041	9,598	18,412	10,789
All	159,343	222,283	3,134K	4,761K	2,349K
All\French	43,672	78,311	481,333	1,193K	937,683

Table 3: Comparison of yields, number of clean document pairs and sentence pairs, produced via URL and CDA on WMT-16 Deep Crawl. Column **CDA\ URL** reports the number of novel sentence pairs exclusively extracted via CDA.

not provide some details of the experiment setting, which are not critical to illustrate our findings.

4.5.1 Robustness Benchmark

We ran the STRAND pipeline end-to-end to extract parallel sentence pairs from document pairs identified by URL and CDA replacing URL. We employ a crawl dataset larger than a typical monthly crawl archive from CommonCrawl. The dataset is also considered densely multilingual. We target six mid-tier languages that are *not* in the top-10 high-resourced languages. It shows that CDA can increase additional 27% English parallel tokens over the selected languages.

Automatic Evaluation We first study the quality of the extracted parallel data, especially the addition of 27% produced by CDA, using automatic MT evaluation. Specifically, for each language, we compare the translation models trained by two equal-sized parallel sentence pairs sampled from the exclusive pairs extracted by URL and CDA individually, i.e., after removing common pairs extracted by both methods. We train vanilla

Lang.	#Dom.	#EN-docs	#XX-docs	$\frac{\|CDA\|}{\|URL\|}$
Turkish	37	1,434,923	71,034	1.77
Czech	69	2,333,914	831,072	7.07
Japanese	84	2,097,664	757,872	1.90
Russian	125	2,918,594	1,163,258	1.09
Italian	239	5,770,684	1,112,868	1.05
Chinese	108	3,061,782	207,211	0.99
All	662	17,617,561	4,143,315	1.24

Table 4: Parallel English tokens extracted by CDA and URL on CommonCrawl Sextet
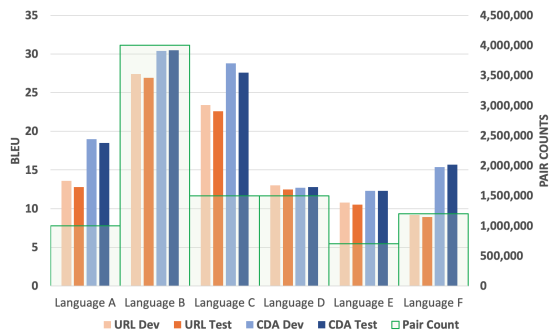

Figure 2: Automatic evaluation of MT models trained by the same amount of pairs sampled from pairs extracted exclusively by either URL or CDA. The primary y-axis (left) indicates the BLEU score, while the secondary y-axis (right) indicates the number of parallel pairs.

`seq2seq` models using Sockeye[10]. We report the MT performance in BLEU scores on our MT evaluation data in Figure 2. The results indicate that the models trained using novel sentence pairs extracted by CDA consistently give better translation models.

Human Evaluation We had linguists manually verify the extracted parallel sentence pairs produced by the pipeline using either URL or CDA. Specifically, we randomly sampled 500 sentence pairs extracted from each pipeline using either URL and CDA for Language A and Language E for human evaluation (we do not remove the common pairs in this evaluation). The selection of these languages is based on their low performance reported during the automatic evaluation in Figure 2. Table 5 shows the result in terms of precision and recall. In general, we find that the quality of pairs produced by both methods is typically comparable. The result also confirms the robustness of our proposed CDA under stress evaluations.

	URL	CDA
Language A	P=96, R=87	P=98, R=89
Language E	P=94 ,R=81	P=92, R=90

Table 5: Human verification for parallel sentence pairs extracted from URL and CDA in precision (P) and recall (R).

4.5.2 Identifying Non-standard Language Identifiers for URL

Even though URL method operates decently fast, it requires language identifiers usually collected manually. This task is challenging as language identifiers used in different web-domains typically do not follow any standard. For example, the language patterns for Czech may include `czech`, `cze`, `ces`, `cz`, `cs`, `cesky`. The result in Table 4 also suggests the limitation of URL for Czech language. In this exercise, we examine URL pairs matched by CDA method to extract relevant language identifiers. Specifically, we focus on URL pairs distanced by *one* token and extract the different tokens as candidates. For example, `en` and `vi_vn` are extracted as candidates for (`www.visitsingapore.com/en/`, `www.visitsingapore.com/vi_vn/`). We curated the candidates, identified additional novel language identifiers, and feeded them to URL method. The exercise helped significantly increase the yields for multiple low-resourced languages at 453%, 295%, and 266%. For example, we found additional identifiers for Chinese language: `chs`, `chn`, `c`, `zho`, `zht`, `cht`, `webcn`, `sc`, `tc`, `chinese_gb`, `chinese_big5`, besides other popular `zh`, `chi`, and `zho`.

4.5.3 Identifying Multilingual Web-domains

We study the application of URL and CDA in identifying densely multilingual web-domains. Specifically, we compare the yield of parallel content, in the total number of extracted parallel English tokens, from two different datasets processed by the same pipeline. The datasets differ in whether their web-domains are identified as multilingual by URL or CDA.

On a sufficiently large dataset, we first ran URL and selected those web-domains having at least 100 candidate pairs. Subsequently, we ran CDA and selected those with at least 100 candidate pairs on the remaining of the dataset, i.e., those not selected by URL. We randomly selected 10,000 domains from

	Domains by URL	Domains by CDA
Parallel Token Count	23.3MM	69.5MM
Host Count	10,000	10,000
Dataset Size (TB)	6	3.1

Table 6: Number of parallel English tokens (in MM) from two deep crawls seeded by URL and CDA.

each group to create the two datasets, namely "Domains by URL" and "Domains by CDA," respectively. We applied the same pipeline using both methods for aligning documents on each dataset. We then computed the yield of parallel English tokens extracted from each setting. Table 6 shows the results. These indicate that CDA can identify densely multilingual web-domains effectively.

In particular, given the same number of web-domains, the dataset identified by CDA can produce almost 3× more parallel data with a size of only half of the dataset identified by URL. This finding suggests that the yield of parallel content from web-domains identified by CDA is 6× higher than those identified by URL. The finding is essential in optimizing the parallel extraction pipeline and identifying better densely multilingual web content.

4.6 Cost Analysis

As presented, we anticipate two cost types when extending CDA to support a new language: (i) building a lexical translation model and (ii) processing more documents. The former is a one-time cost, while the latter is dataset dependent.

Specifically, we have shown in Section 4.2.2 that a lexical translation model can be built using either statistical method IBM-1 with parallel data or neural-based unsupervised method (Lample et al., 2018); we observed comparable performance of CDA when using a model built by these methods. Given the rapid advance in deep neural language models, it is increasingly possible to obtain such resources for low-resourced languages. This suggests that we will be able to leverage recent advances in neural-based NLP to continuously extend CDA for many more languages.

Regarding the execution time, the primary bottleneck typically is due to the scoring of all possible alignments between English and non-English documents. Even though this scoring step is quadratic, this workload is perfectly parallel. With proper en-

gineering optimization, we empirically found out that it is possible to bring the run-time for CDA to be within 2.5× than the one of URL's for 20 low-resourced languages and on a sufficiently large dataset. This optimized cost is crucial in enabling a spectrum of multilingual applications, including cross-lingual information retrieval and enabling MT services for scarce languages.

5 Conclusion

We presented our content-based document alignment for web data, CDA, which projects multilingual documents to a common space for similarity calculation. We also described our effort to collect and create benchmark datasets to study different performance aspects of the proposed method. The results show that CDA is efficient when projecting multilingual documents in one go for as many as 28 languages. Moreover, we also explain the different types of benchmarking for CDA under industrial settings.

The results show that our proposed method is robust when processing huge datasets and useful in identifying non-standard language identifiers and multilingual web-domains. Finally, and most importantly, the only significant resource required by CDA is the lexical translation dictionary: this can be easily built thanks to the recent advance in learning multilingual embeddings.

Future applications of CDA can be many. For example, the URLs paired with CDA can be used to improve the coverage for URL-based methods (e.g., Czech case in Table 4) and to study the web structure of multilingual content. Moreover, the robustness and extensibility of CDA make it applicable to other multilingual processing systems, including cross-lingual search and retrieval.

Acknowledgements

References

Peter F. Brown, Stephen Della-Pietra, Vincent Della-Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation. *Comp.Ling.*

Christian Buck and Philipp Koehn. 2016a. Findings of the wmt 2016 bilingual document alignment shared task. In *WMT 2016*, pages 554–563, Berlin, Germany.

Christian Buck and Philipp Koehn. 2016b. Quick and Reliable Document Alignment via TF/IDF-weighted Cosine Distance. In *WMT 2016*, pages 672–678, Berlin, Germany.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *EMNLP: System Demonstrations*, pages 169–174, Brussels, Belgium.

Aswarth Abhilash Dara and Yiu-Chang Lin. 2016. YODA system for WMT16 shared task: Bilingual document alignment. In *WMT 2016*, Berlin, Germany.

Ahmed El-Kishky and Francisco Guzmán. 2020. Massively multilingual document alignment with cross-lingual sentence-mover's distance. In *AACL-IJCNLP 2020*, pages 616–625, Suzhou, China.

Luís Gomes and Gabriel Pereira Lopes. 2016. First steps towards coverage-based document alignment. In *WMT 2016*, pages 697–702, Berlin, Germany.

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 26–31, Marseille, France. European Language Resources Association.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit 2005*, pages 79–86, Phuket, Thailand.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *ICLR 2018*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*, sixteenth edition. SIL International, Dallas, TX, USA.

Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL 2004*, pages 265–272, Boston, Massachusetts, USA.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Santanu Pal, Partha Pakray, and Sudip Kumar Naskar. 2014. Automatic building and using parallel resources for SMT from comparable corpora. In *HyTra*.

Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Emilia Käsper, and Irina Temnikova. 2004. Multilingual and cross-lingual news topic tracking. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 959–965, Geneva, Switzerland. COLING.

Philip Resnik. 1999. Mining the web for bilingual text. In *ACL 1999*, pages 527–534, College Park, Maryland, USA.

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

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *WMT 2018*, Brussels, Belgium.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *ACL 2013*, pages 1374–1383, Sofia, Bulgaria.

Ralf Steinberger, Bruno Pouliquen, and Johan Hagman. 2002. Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC. In *Computational Linguistics and Intelligent Text Processing, Third International Conference, CICLing'2002*, number 2276 in Lecture Notes in Computer Science, LNCS, pages 415–424. Springer-Verlag.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *COLING 2010*, China.

Thuy Vu, AiTi Aw, and Min Zhang. 2009. Feature-based method for document alignment in comparable news corpora. In *EACL 2009*, pages 843–851, Athens, Greece.

Ivan Vulic and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. pages 363–372. ACM; New York, NY.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *EMNLP 2017*, Denmark.