Findings of the WMT 2016 Bilingual Document Alignment Shared Task

Christian Buck
School of Informatics
University of Edinburgh
Scotland, European Union

Philipp Koehn
Johns Hopkins University
phi@jhu.edu
and
School of Informatics
University of Edinburgh
Scotland, European Union

Abstract

This paper presents the results of the WMT16 Bilingual Document Alignment Shared Task. Given crawls of web sites, we asked participants to align documents that are translations of each other. 11 research groups submitted 19 systems, with a top performance of 95.0%.

1 Introduction

Parallel corpora are especially important for training statistical machine translation systems, but so far the collection of such data within the academic research community has been ad hoc and limited in scale. To promote this research problem we organized a shared task on one of the core processing steps in acquiring parallel corpora from the web: aligning bilingual documents from crawled web sites.

The task is to identify pairs of English and French documents from a given collection of documents such that one document is the translation of the other. As possible pairs we consider all pairs of documents from the same webdomain for which the source side has been identified as (mostly) English and the target side as (mostly) French.

Lack of data in some cases has held back research. To give an example, there are significant research efforts on various Indic languages (Post et al., 2012; Joshi et al., 2013; Singh, 2013), but this work has been severely hampered, since it uses very small amounts of data. But even for the language pairs tackled in high profile evaluation campaigns, such as the ones organized around WMT, IWSLT, and even NIST, we use magnitudes of data less than what has been reported to be used in the large-scale efforts of Google or Microsoft. This diminishes the value of research findings: reported improvements for methods may not hold up once more data is used. Work in reduced data settings may also distract from efforts to tackle problems that do not go away with more data, but are inherent limitations of current models.

2 Related Work

Although the idea of crawling the web indiscriminately for parallel data goes back to the 20th century (Resnik, 1999), work in the academic community on extraction of parallel corpora from the web has so far mostly focused on large stashes of multilingual content in homogeneous form, such as the Canadian Hansards, Europarl (Koehn, 2005), the United Nations (Rafalovitch and Dale, 2009; Ziemski et al., 2015), or European Patents (Täger, 2011). A nice collection of the products of these efforts is the OPUS web site[1] (Skadiņš et al., 2014).

These efforts focused on individual web sites allow for writing specific rules for aligning documents as well as extracting and aligning content. Scaling these manual efforts to thousands or millions of web sites is not practical.

A typical processing pipeline breaks up parallel corpus extraction into five steps:

- Identifying web sites with bilingual content
- Crawling web sites
- Document alignment
- Sentence alignment
- Sentence pair filtering

For each of these steps, there has been varying amount of prior work and for some tools are readily available. Since there has been comparatively little work on document alignment, we picked this problem as the subject for the shared task this year, but other steps are valid candidates for future tasks.

[1] http://opus.lingfil.uu.se/

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 554–563,
Berlin, Germany, August 11-12, 2016.

2.1 Web Crawling

Web crawling is a topic that has not received much attention from a specific natural language processing perspective. There are a number of challenges, such as identification of web sites with multilingual content, avoiding to crawl web pages with identical textual content, learning how often to recrawl web sites based on frequency of newly appearing content, avoiding crawling of large sites that have content in different languages that is not parallel, and so on.

We used for the preparation of this shared task the tool Httrack[2] which is a general web crawler that can be configured in various ways. Papavassiliou et al. (2013) present the focused crawler ILSP-FC[3] that integrates crawling more closely with subsequent processing steps like text normalization and deduplication.

2.2 Document Alignment

Document alignment can be defined as a matching task that takes a pair of documents and computes a score that reflects the likelihood that they are translations of each others. Common choices include edit-distance between linearized documents (Resnik and Smith, 2003), cosine distance of idf-weighted bigram vectors (Uszkoreit et al., 2010), and probability of a probabilistic DOM-tree alignment model (Shi et al., 2006).

2.3 Sentence Alignment

The topic of sentence alignment has received a lot of attention, dating back to the early 1990s with the influential Church and Gale algorithm that is language-independent and easy to implement. It relies on relative sentence lengths for alignment decisions and hence is not tolerant to noisy input.

Popular tools are Hunalign[4] (Varga et al., 2005), Gargantua[5] (Braune and Fraser, 2010), Bilingual Sentence Aligner (Moore, 2002) Bleualign[6] (Sennrich and Volk, 2010), and Champollion[7] (Ma, 2006). Shi and Zhou (2008) make use of the HTML structure to guide alignment. All of these use bilingual lexicons which may have to be provided upfront or are learned unsupervised.

It is not clear, which of these tools fares best with noisy parallel text that we can expect from web crawls, which may have spurious content and misleading boilerplate.

2.4 Filtering

A final stage of the processing pipeline filters out bad sentence pairs. These exist either because the original web site did not have any actual parallel data (garbage in, garbage out), or due to failures of earlier processing steps.

As Rarrick et al. (2011) point out, a key problem for parallel corpora extracted from the web is filtering out translations that have been created by machine translation. Venugopal et al. (2011) propose a method to watermark the output of machine translation systems to aid this distinction. Antonova and Misyurev (2011) report that rule-based machine translation output can be detected due to certain word choices, and machine translation output due to lack of reordering.

This year, a shared task on sentence pair filtering[8] was organized, albeit in the context of cleaning translation memories which tend to be cleaner that the data at the end of a pipeline that starts with web crawls.

2.5 Comprehensive Tools

For a few language pairs, there have been individual efforts to cast a wider net, such as the billion word French–English corpus collected by Callison-Burch et al. (2009), or a 200 million word Czech–English corpus collected by Bojar et al. (2010). Smith et al. (2013) present a set of fairly basic tools to extract parallel data from the publicly available web crawl CommonCrawl[9].

In all these cases, the corpus collection effort reinvented the wheel and wrote dedicated scripts to download web pages, extract text, and align sentences, with hardly any description of the methods used.

Our data preparation for the shared task builds partly on Bitextor[10], which is a comprehensive pipeline from corpus crawling to sentence pair cleaning (Esplà-Gomis, 2009).

[2] https://www.httrack.com/
[3] http://nlp.ilsp.gr/redmine/projects/ilsp-fc
[4] http://mokk.bme.hu/en/resources/hunalign/
[5] https://sourceforge.net/projects/gargantua/
[6] https://github.com/rsennrich/Bleualign
[7] https://sourceforge.net/projects/champollion/

[8] NLP4TM 2016: Shared task http://rgcl.wlv.ac.uk/nlp4tm2016/shared-task/
[9] http://commoncrawl.org/
[10] https://sourceforge.net/p/bitextor/wiki/Home/

3 Training and Test Data

We made available crawls of web sites (defined as pages under the same webdomain) that have translated content. We also annotated some document pairs to provide supervised training data to the participants of the shared task.

3.1 Terminology

A quick note on terminology: Unfortunately, the notion of *domain* is ambiguous in NLP applications, and we use an unusual meaning of the word in this report. To avoid confusion we will instead use the term webdomain to refer to content from a specific website, e.g,"This page is from the `statmt.org` webdomain." We distinguish between webdomains using their Fully Qualified Domain Name (FQDN). Thus, `www.example.com` and `example.com` are considered to be different webdomains.

We will use *source* to denote English pages and *target* for French ones. This does not imply that translation was performed in that direction. In fact we cannot know if translation from one side to the other was performed at all, both sides could possibly be translations of a third language document.

The task was organized as part of the First Conference on Machine Translation (WMT), and all data can be downloaded from its web page[11].

3.2 Data Preparation

We crawled full web sites with the web site copyer **HTTrack**, from the homepage down, restricted to HTML content. Web sites differed significantly in their size, from a few hundred pages to almost 100,000.

In the test data we removed all duplicates from the crawl[12]. Duplicates are defined as web pages, whose text content is identical. Duplicates may differ in markup and URL. To extract the text we used a Python implementation of the HTML5 parser to extract text as a browser would see it. As the text is free of formatting, determining whitespace is important. While generally following the standard, e.g. inserting line breaks after block level elements[13], we found that inserting spaces around `<span>` tags helps tokenization as these are often visually separated using CSS.

We restricted the task to the alignment of French and English documents, so we filtered out all web pages that are not in these two languages. However, we did not expect that participants would develop language-specific approaches. To detect the language of a document we feed the extracted text into an automatic language detector [14]. We note that language detection is a noisy process and many pages contain mixed language context, for example English boilerplate but French content. We take the overall majority language per page as the document language.

We decided to have a large collection of web sites, to encourage methods that can cope with various types of web sites, such as differing in size, balance in the number of French and English pages, and so on.

Given the large number of correct document pairs, we did not even attempt to annotate all of them, but instead randomly selected a subset of pages and identified their corresponding translated page. We augmented this effort with aligned document pairs that are indicated at the web site **Linguee**[16], a searchable collection of parallel corpora, in which each retrieved sentence is annotated with its source web page.

The task then is to find these document pairs. Since this is essentially a recall measure, which can be gamed by returning all possible document pairs, we enforce a 1-1 rule, so that participants may align each web page only once.

3.3 Training Data

As training data we provide a set of 1,624 EN-FR pairs from 49 webdomains. The number of annotated document pairs per webdomain varies between 4 and over 200. All pairs are from within a single webdomain, possible matches between two different webdomains, e.g. `siemens.de` and `siemens.com`, are not considered in this task.

The full list of webdomains in the training data is listed in Table 1. Webdomains range in size from 33×29 pages (`schackportalen.nu`) to 24,325×43,045 pages (`www.nauticnews.com`).

3.4 Test Data

For testing, we provide 203 additional crawls of new webdomains, distinct from the ones in the training data in the same format. No aligned pairs

[11] `http://www.statmt.org/wmt16/bilingual-task.html`

[12] Because we provide the extracted texts of the training pages participants were able to do the same

[13] `https://developer.mozilla.org/en-US/docs/Web/HTML/Block-level_elements`

[14] Compact Language Detector 2 (CLD2)[15]

[16] `http://www.linguee.com/`

Website	Source Documents	Target Documents	Possible Pairs	Train Pairs
cineuropa.mobi	23 050	15 972	368 154 600	73
forcesavenir.qc.ca	3 592	3 982	14 303 344	8
galacticchannelings.com	4 231	1 283	5 428 373	9
golftrotter.com	377	361	136 097	8
ironmaidencommentary.com	6 028	635	3 827 780	41
kicktionary.de	2 752	888	2 443 776	29
kustu.com	1 544	1 511	2 332 984	13
manchesterproducts.com	15 621	9 651	150 758 271	10
minelinks.com	736	212	156 032	66
pawpeds.com	983	135	132 705	19
rehazenter.lu	201	317	63 717	16
tsb.gc.ca	5 885	5 828	34 297 780	236
virtualhospice.ca	43 500	22 327	971 224 500	46
www.acted.org	3 333	2 431	8 102 523	21
www.artsvivants.ca	5 487	1 368	7 506 216	12
www.bonnke.net	414	129	53 406	27
www.cyberspaceministry.org	1 534	958	1 469 572	29
www.dfo-mpo.gc.ca	25 277	19 087	482 462 099	97
www.ec.gc.ca	12 266	15 404	188 945 464	26
www.eu2005.lu	5 649	5 704	32 221 896	34
www.inst.at	3 203	543	1 739 229	62
www.krn.org	115	115	13 225	67
www.lameca.org	692	1 567	1 084 364	6
www.pawpeds.com	1 011	136	137 496	43
bugadacargnel.com	919	779	715 901	19
cbsc.ca	1 595	904	1 441 880	20
creationwiki.org	8 417	203	1 708 651	22
eu2007.de	3 201	2 488	7 964 088	11
eu.blizzard.com	10 493	6 640	69 673 520	10
iiz-dvv.de	1 160	894	1 037 040	67
santabarbara-online.com	1 151	1 099	1 264 949	11
schackportalen.nu	33	29	957	14
www.antennas.biz	812	327	265 524	30
www.bugadacargnel.com	919	779	715 901	7
www.cgfmanet.org	9 241	6 260	57 848 660	25
www.dakar.com	17 420	14 582	254 018 440	45
www.eohu.ca	2 277	2 136	4 863 672	4
www.eu2007.de	3 249	2 535	8 236 215	11
www.fao.org	11 931	5 004	59 702 724	6
www.luontoportti.com	3 645	1 796	6 546 420	30
www.nato.int	40 063	8 773	351 472 699	36
www.nauticnews.com	24 325	43 045	1 047 069 625	21
www.prohelvetia.ch	5 209	4 421	23 028 989	7
www.socialwatch.org	13 803	2 419	33 389 457	21
www.summerlea.ca	434	338	146 692	58
www.the-great-adventure.fr	2 038	2 460	5 013 480	18
www.ushmm.org	10 472	967	10 126 424	26
www.usw.ca	5 006	2 247	11 248 482	83
www.vinci.com	3 564	3 374	12 024 936	24
Total	348 858	225 043	4 246 520 775	1 624

Table 1: Training data statistics.

are provided for the any of these domains. We removed exact duplicates of pages, keeping only one instance. Otherwise, we processed the data in the same way as the training data.

3.5 Data Format

The training document pairs are specified as one pair per line:

`Source_URL<TAB>Target_URL`

For the crawled data we provide one file per webdomain in `.lett` format adapted from Bitextor. This is a plain text format with one line per page. Each line consists of 6 tab-separated values:

- Language ID (e.g. en)
- Mime type (always text/html)
- Encoding (always charset=utf-8)
- URL
- HTML in Base64 encoding
- Text in Base64 encoding

To facilitate use of the `.lett` files we provide a simple reader class in Python. We make sure that the language id is reliable, at least for the documents in the train and test pairs.

Text extraction was performed using an HTML5 parser. As the original HTML pages are available, participants are welcome to implement their own text extraction, for example to remove boilerplate.

Additionally, we have identified spans of French text in French documents for which we produced English translations using MT. We use a basic Moses statistical machine translation engine (Koehn et al., 2007) trained on Europarl and News Commentary with decoding settings geared towards speed (no lexicalized reordering model, no additional language model, cube pruning with pop limit 500).

These translations are not part of the lett files but provided separately. The format for the source segments and target segments is

`URL<TAB>Text`

where the same URL might occur multiple times if several lines/spans of French text were found. The URLs can be used to identify the corresponding documents in the .lett files.

3.6 Baseline Method

We provide a baseline systems that relies on the URL matching heuristic used by Smith et al. (2013). Here two URLs are considered a pair if both can be transformed into the same string through stripping of language identifiers. Strings indicating languages are found by splitting a large number of randomly sampled URLs into components and manually picking substrings that correlate with the detected language.

We further improve the approach by allowing matches where only one URL contains a strip-able language identifier, e.g. we match `x.com/index.htm` and `x.com/fr_index.htm`. If a URL has several matching candidates we pick the one that requires the fewest rewrites, i.e. we prefer the pair above over `x.com/en/index.htm` `x.com/fr_index.htm`.

The baseline achieves roughly 60% recall, compared to 95.0% of the best submission.

4 Evaluation

Our main evaluation metric is recall of the known pairs, i.e. what percentage of the aligned pages in the test set are found. We strictly enforce the rule that every page may only be aligned once, so that participants cannot just align everything. After a URL has been seen as part of a submitted pair, all later occurrences are ignored.

After we released the gold standard alignments, a number of participants pointed out that some predicted document pairs were unfairly counted as wrong, even if their content differed only insignificantly from the gold standard.

To give an example, the web pages

`www.taize.fr/fr_article10921.html?chooselang=1`

and

`www.taize.fr/fr_article10921.html`

are almost identical, but the first offers a checkbox to select a language, while the second does not. Since the text on the pages differs slightly, these were not detected as (exact) duplicates.

To address this problem, we also included a **soft scoring metric** which counts such near-matches as correct. We chose that to be a close duplicate, the edit distance between the text of two pages, normalized by the maximum of their lengths (in characters) must not exceed 5%.

If we observe a predicted pair (s, t) that is not in the gold set, but (s, t') is and $\mathrm{dist}(t, t') \leq 5\%$, then this pair is still counted as correct. The same applies for a close duplicate s' of s but not both as we still follow the 1-1 rule.

Acronym	Participant
ADAPT	ADAPT Research Center, Ireland (Lohar et al., 2016)
BADLUC	University of Montréal, Canada (Jakubina and Langlais, 2016)
DOCAL	Vicomtech (Azpeitia and Etchegoyhen, 2016)
ILSP/ARC	Athena Research and Innovation Center, Greece (Papavassiliou et al., 2016)
JIS	JIS College of Engineering, Kalyani, India (Mahata et al., 2016)
MEVED	Lexical Computing / Masaryk University, Slovakia (Medved et al., 2016)
NOVALINCS	Universidade Nova de Lisboa, Portugal (Gomes and Pereira Lopes, 2016)
UA PROMPSIT	University of Alicante / Prompsit: Bitextor, Spain (Esplà-Gomis et al., 2016)
UEDIN COSINE	University of Edinburgh, Scotland — Buck (Buck and Koehn, 2016)
UEDIN LSI	University of Edinburgh, Scotland — German (Germann, 2016)
UFAL	Charles University in Prague, Czech Republic (Le et al., 2016)
YSDA	Yandex School of Data Analysis, Russia (Shchukin et al., 2016)
YODA	Carnegie Mellon University (Dara and Lin, 2016)

Table 2: List of participants

5 Results

11 research groups participated in the shared task, some with multiple submissions. The list of participants is shown in Table 2, with a citation of their system descriptions, which are included in these conference proceedings.

Each participant submitted one or more collections of document pairs. We enforced the 1-1 rule on the collections, and scored them against the gold standard. Results are summarized in Table 3. Almost all systems outperformed the baseline by a wide margin. The best system is NOVALINCS-URL-COVERAGE with 2,281 correct pairs, 95.0% of the total.

Note that the submissions varied in the number of document pairs, but after enforcing the 1-1 rule, most submissions comprise about 200,000-300,000 document pairs.

Table 4 displays the results with soft scoring. Essentially, every system improved, mostly by around 3%. The top two performers swapped places, with YODA now having the best showing with 96.0%. We also experimented with a tighter threshold of 1% which gave almost identical results.

6 System Descriptions

NOVALINCS (Gomes and Pereira Lopes, 2016) submitted 3 systems that use a phrase table from a phrase-based statistical machine translation system to compute coverage scores, based on the ratio of phrase pairs covered by a document pair. In addition to the purely coverage-based system, NOVALINCS-COVERAGE (88.6%), they also submit a system that uses coverage-based matching as a preference over URL matching NOVALINCS-COVERAGE-URL (85.8%) and the converse system that prefers URL matching over coverage-based matching NOVALINCS-URL-COVERAGE (95.0%).

YODA (Dara and Lin, 2016) submitted one system (93.9%) that uses the machine translation of the French document, and finds the English corresponding document based on bigram and 5-gram matches, assisted by a heuristics based on document length ratio.

UEDIN1 (Buck and Koehn, 2016) submitted one system (89.1%) that uses cosine similarity between tf/idf weighted vectors, extracted by collecting n-grams from the English and machine translated French text. They compare many hyperparameters such as weighting schemes and two pair selection algorithms.

DOCAL (Azpeitia and Etchegoyhen, 2016) submitted one system (88.6%) that used word translation lexicons to compute document similarity scores based on bag-of-word representations. They expand a basic translation lexicon by adding all capitalized tokens, numbers, and longest common prefixes of known vocabulary items.

UEDIN2 (Germann, 2016) submitted 2 systems based on word vector space representations of documents using latent semantic indexing and URL matching, UEDIN LSI (85.8%) and UEDIN LSI (87.6%). In addition to a global cosine similarity score, a local similarity score is computed by re-centering the vector around the mean vector for a webdomain.

Name	Predicted pairs	Pairs after 1-1 rule	Found pairs	Recall %
ADAPT	61 094	61 094	644	26.8
ADAPT-V2	69 518	69 518	651	27.1
BADLUC	681 610	263 133	1 905	79.3
DOCAL	191 993	191 993	2 128	88.6
ILSP-ARC-PV42	291 749	287 860	2 040	84.9
JIS	323 929	28 903	48	2.0
MEDVED	155 891	155 891	1 907	79.4
NOVALINCS-COVERAGE-URL	207 022	207 022	2 060	85.8
NOVALINCS-COVERAGE	235 763	235 763	2 129	88.6
NOVALINCS-URL-COVERAGE	235 812	235 812	**2 281**	**95.0**
UA PROMPSIT BITEXTOR 4.1	95 760	95 760	748	31.1
UA PROMPSIT BITEXTOR 5.0	157 682	157 682	2 001	83.3
UEDIN1 COSINE	368 260	368 260	2 140	89.1
UEDIN2 LSI	681 744	271 626	2 062	85.8
UEDIN2 LSI-V2	367 948	367 948	2 105	87.6
UFAL-1	592 337	248 344	1 953	81.3
UFAL-2	574 433	178 038	1 901	79.1
UFAL-3	574 434	207 358	1 938	80.7
UFAL-4	1 080 962	268 105	2 023	84.2
YSDA	277 896	277 896	2 021	84.1
YODA	318 568	318 568	2 256	93.9
Baseline	148 537	148 537	1 436	59.8

Table 3: Official Results of the WMT16 Bilingual Document Alignment Shared Task.

Name	Pairs found	Δ	Recall	Δ	Rank	Δ
ADAPT	726	+82	30.2	+3.4	20	0
ADAPT-V2	733	+82	30.5	+3.4	19	0
BADLUC	2 062	+157	85.9	+6.5	13	+3
DOCAL	2 235	+107	93.1	+4.5	4	+1
ILSP-ARC-PV42	2 185	+145	91.0	+6.0	7	+2
JIS	48	0	2.0	0.0	21	0
MEDVED	1 986	+79	82.7	+3.3	15	0
NOVALINCS-COVERAGE-URL	2 130	+70	88.7	+2.9	9	−1
NOVALINCS-COVERAGE	2 192	+63	91.3	+2.6	6	−2
NOVALINCS-URL-COVERAGE	2 303	+22	95.9	+0.9	2	−1
UA PROMPSIT BITEXTOR 4.1	775	+27	32.3	+1.1	18	0
UA PROMPSIT BITEXTOR 5.0	2 117	+116	88.1	+4.8	10	+2
UEDIN1 COSINE	2 227	+87	92.7	+3.6	5	−2
UEDIN2 LSI	2 146	+84	89.3	+3.5	8	−1
UEDIN2 LSI-V2	2 281	+176	95.0	+7.3	3	+3
UFAL-1	2 060	+107	85.8	+4.5	14	−1
UFAL-2	1 954	+53	81.4	+2.2	17	0
UFAL-3	1 980	+42	82.4	+1.8	16	−2
UFAL-4	2 078	+55	86.5	+2.3	12	−2
YSDA	2 102	+81	87.5	+3.4	11	0
YODA	**2 307**	+51	**96.0**	+2.1	**1**	+1

Table 4: Soft Scoring Results of the WMT16 Bilingual Document Alignment Shared Task, allowing 5% edits between predicted and expected pairing.

ILSP/ARC (Papavassiliou et al., 2016) submitted one system (84.9%), which uses boilerplate removal, and carries out document alignment based on features such as links to documents in the same webdomain, URLs, digits, image filenames and HTML structure. Their paper also describes in detail the open source ILSP Focused Crawler.

YSDA (Shchukin et al., 2016) submitted one system (84.1%) that uses n-gram matches between the machine translation of the French document and the English document. They cluster French and English words into bilingual clusters of up to 90 words, starting with word pairs with high translation probability in both directions, and then adding words that translated well into existing words in a cluster.

UA PROMPSIT (Esplà-Gomis et al., 2016) submitted 2 systems based on Bitextor and describe improvements to the Bitextor toolkit. Their submissions contrast the old version of the tool, UA PROMPSIT BITEXTOR 4.1 (31.1%), with the recent release, UA PROMPSIT BITEXTOR 5.0 (83.3%). Improved document alignment quality is based on various new features: ratio of shared links, similarity of link URLs, ratio of shared images, binary feature indicating if the documents are linked, and similarity of URLs, in addition to the old features bag of words similarity using a translation dictionary and DOM structure similarity.

UFAL (Le et al., 2016) submitted 4 systems, each using a different method. UFAL-1 (81.3%) uses identical word matches by also considering their position in the text. UFAL-2 (79.1%) matches translations of French documents with English documents based on word occurrence probabilities. UFAL-3 (80.7%) adds Levenshtein distance on URLs to this method. UFAL-4 (84.2%) combines UFAL-1 and UFAL-3.

MEDVED (Medved et al., 2016) submitted one system (79.4%), which determines the top 100 keywords based on tf/idf scores for each document and uses word translation dictionaries to match them.

BADLUC (Jakubina and Langlais, 2016) submitted one system (79.3%) that uses the information retrieval tool Apache Lucene to create two indexes, on URLs and text content, and retrieves the most similar documents based on variants of td/idf scores. Both monolingual queries and bilingual queries based on a word translation dictionary are performed.

ADAPT (Lohar et al., 2016) submitted one system (and a revision) that combines similarity metrics computed on ratio of number of sentences in documents, ratio of number of words in the documents, and matched named entities.

JIS (Mahata et al., 2016) submitted one system (2.0%), which uses text matching based on sentence alignment and word dictionaries. Their paper also described improvements over the original submission.

Acknowledgment

This shared task is partially supported by a Google Faculty Research Award. This work was also supported by the European Union's Horizon 2020 research and innovation programme under grant agreement 645487 (MMT).

References

Antonova, A. and Misyurev, A. (2011). Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 136–144, Portland, Oregon. Association for Computational Linguistics.

Azpeitia, A. and Etchegoyhen, T. (2016). Docal - vicomtech's participation in the wmt16 shared task on bilingual document alignment. In *Proceedings of the First Conference on Machine Translation*.

Bojar, O., Liška, A., and Žabokrtský, Z. (2010). Evaluating utility of data sources in a large parallel Czech-English corpus CzEng 0.9. In *Proceedings of LREC2010*.

Braune, F. and Fraser, A. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Coling 2010: Posters*, pages 81–89, Beijing, China. Coling 2010 Organizing Committee.

Buck, C. and Koehn, P. (2016). Quick and reliable document alignment via tf/idf-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation*.

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28,

Athens, Greece. Association for Computational Linguistics.

Dara, A. A. and Lin, Y.-C. (2016). Yoda system for wmt16 shared task: Bilingual document alignment. In *Proceedings of the First Conference on Machine Translation*.

Esplà-Gomis, M. (2009). Bitextor: a free/open-source software to harvest translation memories from multilingual websites. In *MT Summit Workshop on New Tools for Translators*. International Association for Machine Translation.

Esplà-Gomis, M., Forcada, M., Ortiz Rojas, S., and Ferrández-Tordera, J. (2016). Bitextor's participation in wmt'16: shared task on document alignment. In *Proceedings of the First Conference on Machine Translation*.

Germann, U. (2016). Bilingual document alignment with latent semantic indexing. In *Proceedings of the First Conference on Machine Translation*.

Gomes, L. and Pereira Lopes, G. (2016). First steps towards coverage-based document alignment. In *Proceedings of the First Conference on Machine Translation*.

Jakubina, L. and Langlais, P. (2016). Bad luc@wmt 2016: a bilingual document alignment platform based on lucene. In *Proceedings of the First Conference on Machine Translation*.

Joshi, A., Popat, K., Gautam, S., and Bhattacharyya, P. (2013). Making headlines in hindi: Automatic english to hindi news headline translation. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, pages 21–24, Nagoya, Japan. Asian Federation of Natural Language Processing.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Le, T., Vu, H. T., Oberländer, J., and Bojar, O. (2016). Using term position similarity and language modeling for bilingual document alignment. In *Proceedings of the First Conference on Machine Translation*.

Lohar, P., Afli, H., Liu, C.-H., and Way, A. (2016). The adapt bilingual document alignment system at wmt16. In *Proceedings of the First Conference on Machine Translation*.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In *International Conference on Language Resources and Evaluation (LREC)*.

Mahata, S., Das, D., and Pal, S. (2016). Wmt2016: A hybrid approach to bilingual document alignment. In *Proceedings of the First Conference on Machine Translation*.

Medved, M., Jakubícek, M., and Kovář, V. (2016). English-french document alignment based on keywords and statistical translation. In *Proceedings of the First Conference on Machine Translation*.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In Richardson, S. D., editor, *Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002 Tiburon, CA, USA, October 6-12, 2002, Proceedings*, volume 2499 of *Lecture Notes in Computer Science*. Springer.

Papavassiliou, V., Prokopidis, P., and Piperidis, S. (2016). The ilsp/arc submission to the wmt 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation*.

Papavassiliou, V., Prokopidis, P., and Thurmair, G. (2013). 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, Sofia, Bulgaria. Association for Computational Linguistics.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 154–162, Mon-

treal, Canada. Association for Computational Linguistics.

Rafalovitch, A. and Dale, R. (2009). United Nations General Assembly resolutions: A six-language parallel corpus. In *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*. International Association for Machine Translation.

Rarrick, S., Quirk, C., and Lewis, W. (2011). MT detection in web-scraped parallel corpora. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, pages 422–430. International Association for Machine Translation.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL)*.

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

Sennrich, R. and Volk, M. (2010). MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*.

Shchukin, V., Khristich, D., and Galinskaya, I. (2016). Word clustering approach to bilingual document alignment (wmt 2016 shared task). In *Proceedings of the First Conference on Machine Translation*.

Shi, L., Niu, C., Zhou, M., and Gao, J. (2006). 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.

Shi, L. and Zhou, M. (2008). Improved sentence alignment on parallel web pages using a stochastic tree alignment model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 505–513, Honolulu, Hawaii. Association for Computational Linguistics.

Singh, T. D. (2013). Taste of two different flavours: Which manipuri script works better for english-manipuri language pair smt systems? In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 11–18, Atlanta, Georgia. Association for Computational Linguistics.

Skadiņš, R., Tiedemann, J., Rozis, R., and Deksne, D. (2014). Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. Association for Computational Linguistics.

Täger, W. (2011). The sentence-aligned european patent corpus. In Forcada, M. L., Depraetere, H., and Vandeghinste, V., editors, *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 177–184.

Uszkoreit, J., Ponte, J., Popat, A., and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1101–1109, Beijing, China. Coling 2010 Organizing Committee.

Varga, D., Halaácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005 Conference*, pages 590–596.

Venugopal, A., Uszkoreit, J., Talbot, D., Och, F., and Ganitkevitch, J. (2011). Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1363–1372, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2015). The united nations parallel corpus v1.0. In *International Conference on Language Resources and Evaluation (LREC)*.