YODA System for WMT16 Shared Task: Bilingual Document Alignment

Aswarth Abhilash Dara and **Yiu-Chang Lin**
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, USA
{adara, yiuchanl}@cs.cmu.edu

Abstract

In this paper, we address the task of automatically aligning/detecting the bilingual documents that are translations of each other from a single web-domain as part of WMT 2016. [1] Given the large amounts of data available in each web-domain, a brute force approach like finding similarities between every possible pair is a computationally expensive operation. Therefore, we start with a simple approach on matching just the web page urls after some pre-processing to reduce the number of possible pairings to a small extent. This simple approach obtained a recall of 50% and the exact matches from this approach are removed from further consideration. We built on top of this using an n-gram based approach that uses the partial English translations of French web pages and achieved a recall of 93.71% on the training pairs provided. We also outline an IR-based approach that uses both content and the meta data of each web page url, thereby obtaining a recall of 56.31%. Our final submission to this shared task using n-gram based approach achieved a recall of 93.92%.

1 Introduction

Statistical Machine Translation systems rely a lot on the availability of parallel corpora and the automatic collection of such data so far has been ad hoc and limited in scale. In this paper, we would like to tackle the problem of aligning bilingual documents from crawled websites which is presented as one of the new shared tasks introduced at WMT 2016 i.e., the task of identifying pairs of English and French documents from a given collection of documents such that one document is a translation of another. For each web-domain, we consider all the possible pairs for which the source side has been identified as English and the target side as French. 1,624 EN-FR pairs from 49 web-domains are provided as training data. The number of pairs per web-domain varies between 4 and over 200. All pairs are from within a single web-domain and possible matches between two different web-domains e.g. siemens.de and siemens.com are not considered in this task.

Mirrors of all the web pages in each domain which were crawled using httrack are provided. Each page has the following information: Language ID (e.g. `en`), Mime type (always `text/html`), Encoding (always `charset=utf-8`), URL, HTML in Base64 encoding and Text in Base64 encoding. Additionally, the English translations of French pages using MT for identified spans of text were produced by the organizers. However, it doesn't imply that we have full translations for each and every French web page. In other words, we only have partial translations for a random subset of French web pages.

Table 1 shows the various statistics of the training data set. Among 49 web-domains, *www.nauticnews.com* has the most possible pairings, 1,047,069,625, while *schackportalen.nu* has the least ones, 957. Each web-domain has roughly 87 million pairs on average.

The rest of the paper is organized as follows: Section 2 discusses the various challenges involved in this task. Section 3 gives an overview of the related work happened. The methodology and implementation details are provided in Section 4. Section 5 covers evaluation, results and analysis of the errors on the training data set given. Section 6 concludes the paper with possible future directions

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

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 679–684,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

Web-Domain	Source Pages	Target Pages	Possible Pairings	Train Pairs Provided
www.nauticnews.com	24,325	43,045	1,047,069,625	21
schackportalen.nu	33	29	957	14
Average	7,119	4,592	86,663,689	33

Table 1: Various statistics on the training data. The first row shows the web-domain with most possible pairs and the second row shows the web-domain with the least possible pairs. The last row is the average statistics across 49 domains.

for this work.

2 Challenges

There are various challenges involved in dealing with the bilingual document alignment task and are as follows.

- The primary challenge is that the space of possible pairings is so huge that it is almost impossible to use any brute force approach for comparing every two documents from the source and target with any similarity metric. As shown in Table 1, the largest number of possible pairs come from a single domain in the training set is 1,047,069,625 whereas in the testing set, it comes from cinedoc.org which contains 2,444,607,480 pairs (around 2.4 billion pairs).

- Another challenge involves in obtaining the full translations on either source or target side (even if we restrict ourselves to first n-words in each page) is an expensive operation given the large amounts of processing data.

- Another approach in training the domain specific MT model for each of the 49 web-domains given only 1,624 training pairs is not encouraging because of the less availability of training data in each domain. Even if we train, it will not provide any advantage because the testing set web-domains are completely distinct from the training set.

- In addition to this, even if partial translations of the documents on target side are provided, making the most use of them is a crucial issue.

- Furthermore, documents vary in length and no positional information of these translations provided are available.

3 Related Work

In general, most statistical parallel corpus alignment works have focused on the sentence and vocabulary level. Kay and Röscheisen (1993) proposed to align texts with their translations that is based only on internal evidence. The idea of iterating the process of sentence level alignment with the results of vocabulary level alignment reinforce the certainty of both. More specifically, it exploits a partial alignment of the word level to induce a maximum likelihood alignment of the sentence level, which is in turn used in the next iteration, to refine the word level estimate. The algorithm appears to converge to the correct sentence alignment in only a few iterations.

Gale and Church (1991) focused their attention on robust statistics that tends to keep errors of commission low. They introduced a measurement of association between a pair of words based on a two by two contingency table and obtained bilingual vocabularies by presenting the co-occurrence statistics. Melamed (1999) used advanced bi-text mapping by formulating the problem in terms of pattern recognition where the success of a bi-text mapping algorithm lies in how well it performs in these three tasks: signal generation, noise filtering, and search. The proposed Smooth Injective Map Recognizer (SIMR) algorithm integrates innovative approaches to each of these tasks.

There are also works focusing on combining information from both sentence and vocabulary alignments (Moore, 2002), which combined Sentence length based methods and Word correspondence based methods for aligning sentences with their translations in a parallel bilingual corpus. It achieved a high accuracy at a modest computational cost, and required no knowledge of the languages or the corpus beyond division into words and sentences. Nazar (2011) presented a language independent algorithm for the alignment of parallel corpora at the document, sentence ad vocabulary levels. The process consists of the follow-

ing phases: aligning documents with their corresponding translations, aligning sentences inside each pair of selected documents and finally, generating a bilingual vocabulary. For large scale document level alignment, Uszkoreit et al., (2010) proposed a distributed system that reliably mines parallel text from large corpora. In contrast to other approaches which require specific meta data, the system uses only the textual information. In this paper, we take inspiration from this approach and add some interesting heuristics on top of it to obtain a good recall.

Another family of work is to learn an intermediate document representation between documents from the source and target side where similar intermediate concepts are closely projected. There are various kinds of such deep learning models, for example, Deep Structured Semantic Models (Huang et al., 2013), Deep Boltzmann Machines (Salakhutdinov and Hinton, 2009), Stacked Denoising Autoencoder (Vincent et al., 2010), Encoder-Decoder (Cho et al., 2014) and Deep Canonical Component Analysis (Andrew et al., 2013). However, we have not tried any of these deep learning approaches as part of our experiments due to the limited availability of the training data.

4 Methodology

4.1 Baseline

The task organizers provided a baseline approach which uses only the meta data related to url of the web page like stripping the language identifiers etc. and reported the performance on the training data set. The code for the same is available on github.[2] Initially, we thought of building our models on top of this baseline, however the pairs they generate are not exact matches which serves as a main bottleneck in reducing the number of possible pairings.

4.2 Brute Force and URL Patterns Approach

For this task, the training data set consists of 1,624 English-French pairs from 49 web-domains. A straight forward approach is to simply model this as a binary classification problem where 1 indicates that two documents are translations of each other and 0 indicates that they are not. The 1,624 actual training set will become $1,624 \times 1,624$ pseudo training set that can be used to train a skewed classification model. The features can be from the baseline (meta data related to the urls), Machine Translation features (translating the source side and comparing with the possible candidates in the target side using MT evaluation metrics like BLEU (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009), TER (Snover et al., 2006) and also the features can be generated using standard metrics in document-similarity literature). However, due to the quadratic complexity of considering the similarity between every possible pair, we didn't pursue this approach. Also, lack intersection between the training and testing set web-domains add additional problems. Instead, we used a very simple baseline approach based on web page url matching to reduce the space of the number of possible pairings to be considered. We call this approach as URL Patterns.

For every source web page url, we first tokenize this url using the NLTK toolkit [3] and replace the word 'en' (if exists) with 'fr'. We reconstruct the url back and call it as a normalized url. Then, we search for this normalized url in the list of possible target urls and if it exists, we treat them as a possible pair and remove both of them for further consideration. This is a simple approach, however when we evaluated this approach on the training set given, we got around 50% recall. The following approaches are built on top of this simple approach.

4.3 Information Retrieval based Approach

The bilingual document alignment task could be viewed as an Information Retrieval problem where our goal is to retrieve the most relevant document on the source or target side given one from the other side. More specifically, for each French document, we extract queries from its provided English translation and search through all the English documents in the same domain. The one with the highest retrieval score is aligned to that French document. The IR framework implementation is done by the following steps with the use of Whoosh library.[4] First we built indices for every English web page. Second, for each French web page, the query generator extracts all possible trigrams from every sentence but remove those con-

[2] https://github.com/christianbuck/wmt16-document-alignment-task.git

[3] http://www.nltk.org/

[4] https://whoosh.readthedocs.io/en/latest/

taining punctuation, numbers or stop words. Finally, re-ranking is performed by comparing the degree of difference between the tokenized target url and the tokenized source urls. After the re-ranking, we pick the top-most result as the possible candidate and we treat this as the output for that source url.

4.4 N-gram based approach

In this approach, we used the partial English translations provided by task organizers for French web pages.[5] For this particular approach, without loss of generality, we consider French as the source and English as the target. The approach has been inspired from the approach mentioned in (Uszkoreit et al., 2010) and some of the best settings are borrowed with a different set of heuristics. As we are using partial English translations for French web pages and the target is in English, now we have both source and target sides in English language. Two types of indexes are built i.e., forward index and inverted index. Inverted index is built for both bi-grams and 5-grams where the key is either a bi-gram or a 5-gram and the posting list contains web page urls of both English and French web pages. A forward index is built where key is the web page url and the posting list contains the list of bi-grams [6] present in that web page url. This forward index is the one that is used for scoring the similarity between a source and target language web page pair.

The inverted index built based on 5-gram is used for generating the list of possible candidate pairs between both source and target side. Before the generation, we use some heuristics to remove some of the 5-grams that are to be considered. If the size of the posting list of any 5-gram is one, it means that this particular 5-gram is present in only one language and we can safely disregard them from consideration. We also remove the frequent 5-grams for consideration if the posting list of any 5-gram exceeds a certain threshold. Empirically, we found that threshold is 0.1. Similarly, using the inverted index for bi-grams, we propagate the document frequency of a bi-gram into the forward index thereby calculating the inverse document frequency (idf) for this bi-gram. Now, in the forward index, for each web page, we have the list of bi-grams along with the *idf* of each bi-gram which can be used for scoring the cosine similarity between a pair of source and target web page. After this, for each 5-gram, we split the list into French and English web pages and form every possible pair as the candidates. The size of the possible candidate pairs obtained is still large and therefore we use another heuristic to reduce the computational space i.e., document length ratio. Given that the size of the English and French documents won't differ too much, we removed the candidate pairs if the ratio of the document length between the original French and English web pages is less than 0.5. [7] After applying all these, We noticed that the size of the possible candidates generated for all French web pages in each domain is around 1% of the all the possible pairings considered initially. This is a significant reduction in the number of possible pairings that are to be considered thus making the approach computationally feasible.

Finally, given a list of possible English web pages for each French web page, we compute the cosine similarity between the French web page and English web page. If we simply pick the maximum one for each French web page, then there exists a possibility that the English web page we pick may occur subsequently for some other French web page. However, there is a strict constraint enforced by the task organizers that each source web page can be aligned to only target web page and each target web page can get aligned only once. In order to enforce this constraint, we use a simple greedy approach where we first compute the cosine similarities between each French web page and list of possible candidates. Then, we pick the maximum scored pair out of all the possible pairs, and output it and remove it from further consideration. We repeat the same process until all the source side web pages are processed. We submitted our results on the testing data set using this approach to the shared task after conducting various experiments.

5 Evaluation and Results

The evaluation for this problem has been well defined in terms of recall as part of the shared task i.e., what percentage of the test-set pairs are found

[5] Since the translations of text spans are provided, we simply concatenate all partial text spans and translations in the same order provided for each web page before computing the n-grams.

[6] Empirically determined by experimenting with different n-grams on a subset of training data set provided.

[7] We empirically arrived at this threshold by experimenting with different values

Approach	Baseline	URL Patterns	IR-based	n-gram based
Recall	67.92	50.00	56.31	**93.71**

Table 2: Recall on the training set pairs using different approaches

on the predicted test pairs after enforcing the 1-1 rule (each source web page will be matched with at most one target web page and later occurrences of the source web pages are excluded from the evaluation). The performance of the model will be tuned on the training data set provided by the shared task organizers. The performance on the training data set (1624 pairs) is listed in Table 2.

As we observe in Table 2, the IR-based approach didn't work well and in fact it performed worse than the baseline provided by the task organizers. On the other hand, the n-gram based approach worked very well with a recall of 93.71%. We found that out of 49 web-domains, we got a recall of 100% in 31 web-domains. We have also performed an error analysis on the incorrect pairs to get a good understanding of the errors produced by the n-gram based approach. Based on our analysis, it has been found that relying mainly alone on the cosine similarity between a pair of possible candidate pairs is not itself alone, and have to do some re-ranking after computing the initial cosine similarity.

One of the interesting observations we made when looking at the errors is sometimes there exists no one-to-one correspondence between the source web page and target web page. This happens if a target web page is split into multiple target web pages and given only the availability of partial translations, aligning the source web page to maximum similar target web page requires some additional information. Since we only have the partial translations, there is no positional information of each n-gram which will be very useful in calculating the similarity metric. Most of these errors can be easily mitigated if we were provided the entire translation of each French web page instead of providing translations only for some text spans. However, obtaining the full translations for each and every web page is computationally intensive.

We submitted our results on the test set to this shared task using the best approach that is based on n-grams. Our system obtained a recall of 93.92%. It would be very interesting to see how these results change once the re-ranking phase is successfully implemented which serves as a promising future direction.

6 Conclusions and Future Directions

In this paper, we tackled the task of automatically aligning/detecting the bilingual (English and French) documents. With a simple approach of matching urls, we obtained around a recall of 50%. Using an n-gram based approach with interesting heuristics on top of it, we got a recall of around 93.71% and 93.92% on the training and testing data sets respectively.

The future directions for this work include systematically looking at where the errors occurred and increase the performance further. The re-ranking phase using word/document embeddings, structure of the the html document and a lot of other information serves as a straightforward extension to this paper. Another possible direction could be given a web page url as an input, how can we translate it effectively and if the translated url exists in the possible candidates, we can safely remove those pairs from further consideration. However, how to effectively tokenize and translate a web page url is still an interesting question to answer.

References

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1247–1255.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine

translation. *Machine Translation*, 23(2-3):105–115, September.

Robert C Moore. 2002. *Fast and accurate sentence alignment of bilingual corpora*. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ruslan Salakhutdinov and Geoffrey E Hinton. 2009. Deep boltzmann machines. In *International conference on artificial intelligence and statistics*, pages 448–455.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Jakob Uszkoreit, Jay M Ponte, Ashok C Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1101–1109. Association for Computational Linguistics.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408.