WMT2016: A Hybrid Approach to Bilingual Document Alignment

Sainik Kumar Mahata[1], Dipankar Das[2], Santanu Pal[3]
[1]JIS College of Engineering, Kalyani, India
[2]Jadavpur University, Kolkata, India
[3]Universität des Saarlandes, Saarbrücken, Germany
sainik.mahata@gmail.com,
ddas@cse.jdvu.ac.in, santanu.pal@uni-saarland.de

Abstract

Large aligned corpora are required for any computer aided translation system to become effective. In this scenario, bilingual document alignment has gained utmost importance in recent days. We attempt a simple yet effective approach to align URLs (Uniform Resource Locator) within two documents in two languages as a part of WMT2016 Bilingual Document Alignment Shared Task. Our approach includes the processing of URLs and their embedded texts, which serves as the main matching criterion. In order to align the text initially, we have used Gale-Church algorithm, dictionary based translation and Cosine Similarity that in turn helps us to achieve better results in the alignment task.

1 Introduction

Bilingual document alignment has gained utmost importance these days [Brown et al.1991, Warwick et al.1990, Gale and Church1991, Kay and Röscheisen1993, Simard et al.1992, Kupiec1993, Matsumoto et al.1993, Dagan et al.1999, Church1993]. Research on calculating similarity of bilingual comparable corpora is attracting more attention [Vu et al.2009, Pal et al.2014, Tan and Pal2014]. Growth of monolingual data in different languages has made the task for aligning documents difficult [Jagarlamudi et al.2011]. What makes the problem more critical is the fact that one sentence in one language can correspond to many sentences in a different language [Wu1994]. For any translation system to work correctly and efficiently, it has to be fed with a large parallel corpora. Such corpora are very hard to find [Smith et al.2010] since it involves serious manual labour and cost. To eliminate the high cost, computer aided sentence alignment of two different corpora has become very desirable. The presence of such computer aided aligned corpora aids in many Natural Language Processing (NLP) tasks such as Machine Translation, Word Sense Disambiguation as well as Cross Lingual Information Extraction [Patry and Langlais2005].

In our current task, we have worked on the data provided by WMT shared task[1], which had web crawls of 203 websites and were extracted in both English and French. The task was to extract 1-1 pairings of English and French URLs that has the same content but in respective languages. The data contained URLs followed by the text in each of the URLs. The task was to extract the text from both the English and French URLs and align them using our alignment algorithm. After alignment of the text, the URLs to which the text belongs to were also aligned. Our algorithm makes use of concepts given by [Gale and Church1991], translation of words using a dictionary created by Anymalign package [Lardilleux et al.2012] and the concept of Cosine Similarity. The following section will document the algorithm. Working of the algorithm will be shown in Section 2, followed by the results in Section 3.

2 Proposed System

2.1 Text and URL Extraction

The given *.lett* files are opened and the URL as well as the texts are extracted. The extracted text and URLs are given IDs so that it becomes easy to align the URLs after aligning the texts. The process is shown in Figure 1.

2.2 Text Selection using Algorithm proposed by Gale-Church

In their paper [Gale and Church1991], Gale and Church suggested that the source sentence and its

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

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 724–727,
Berlin, Germany, August 11-12, 2016.

```
en  text/html   charset=utf-8   http://academiedesprez.org/mailgb.php (1)   Message to Gilbert Blin (1)
en  text/html   charset=utf-8   http://academiedesprez.org/hamfelt/index.htm (2)    Message for / pour Gilbert Blin (2)
fr  text/html   charset=utf-8   http://academiedesprez.org/ (1) Académie Desprez (1)
en  text/html   charset=utf-8   http://academiedesprez.org/eng/dubos1eng.htm  (3)   Message from / Message de : (3)
en  text/html   charset=utf-8   http://academiedesprez.org/eng/applications9009eng.htm (4)  Name / Nom (4)
en  text/html   charset=utf-8   http://academiedesprez.org/eng/communication1eng.htm  (5)   Firstname / Prénom (5)
fr  text/html   charset=utf-8   http://academiedesprez.org/eng/musicales5eng.htm (6)    Présenté par Gilbert Blin (6)
```

Figure 1: Text and URL extraction

```
Message to Gilbert Blin A cadémie D esprez  Séjour de Lena Dahlström    17h Accueil à Paris Où ? Musée Cognacq-Jay  20h Dîner pour
Lena 10h Musée du Louvre Où ? Cinéma Saint-Lambert   Où ? Marc Anselmi   9h Château de Versailles    Où ? Maison Malbranche  21h
Dîner pour Lena 10h Musée de l'Armée    Où ? Musée Galliera Métro: Goncourt ou Belleville   Où ? Marché Saint-Pierre    Où ?
Maison Lesage  13 rue Grange Batelière 20h Dîner pour Lena Métro: Porte de Clignancourt    Conservatrice au Musée Galliera
Architecture théâtrale en France    Un Théâtre de Voltaire  Houdon, sculpteur des Lumières  La Danse de Mort    Salles, Scènes &
Salons 19h30Théâtre des Champs Elysées Descriptif prévisionnel du projet   Lauréat : Olivier Till  Olympie - 1762 (II-3).  Textes
descriptifs des costumes.    Dossier de Léna Dahlström   Phase 4 - Leyde Budget " costumes " Contrat avec le loueur  Préparation
(essayages et repassage)    9- Fashion in Hair. Images de l'Album Ziesemis. 11- History of Theatre. Musical studies Des Orages
Poetry, declamation & music Le Carnaval de Venise   Présenté par Camille Tanguy Présenté par Gilbert Blin   Présenté par Rémy-
Michel Trotier    Présenté par Gilbert Blin   Séjour de Christer Nilsson  Bourse de Voyage "Servandoni"   Madame Danielle Durand,
Comédie-Française   Samedi 17 Novembre 2001 Dimanche 18 Novembre 2001   Où ? Musée Cognacq-Jay  02330 Condé en Brie Mardi 20
Novembre 2001  10h: les Grands Appartements    20h30 Concert: Trio Wanderer    Où ? Salle Gaveau   Mercredi 21 Novembre 2001
Accompagnateur : Magnus Johansson   Où ? Opéra Bastille Jeudi 22 Novembre 2001  12h30 Eglise Saint Sulpice  Vendredi 23 Novembre
2001   Où ? Palais Garnier Où ? Opéra Bastille Samedi 24 Novembre 2001 Métro: Porte de Clignancourt    Dimanche 25 Novembre 2001
   Qui était Servandoni ?  Fables de La Fontaine   14h Centre Culturel Suédois 11h Château de Vaux-le-Vicomte  10h Bibliothèque-
musée de l'Opéra   21h30 Théâtre du Lucernaire Le Barbier de Séville   9h Musée du Louvre  Déléguée à l'activité commerciale
Monsieur Patrice de Vogüé   Conservateur-archiviste de la Comédie-Française Discover the Armfelt grant: Who was Armfelt ?   Séjour
```

Figure 2: Text Selection according to size, using algorithm proposed by Gale and Church

translated sentence have the same length.

This idea forms the basis of our proposed system. We have found out the length of the source English sentence, that have been extracted from a URL pair, and have found matches in all the target French sentences, extracted from the same URL pair. This results in one-to-many relationship between the English and French sentences. The variance in this step is kept as 2, which means if the length of the French sentences exceeds or falls behind the length of the English sentence by a difference 2, when compared to the source English sentence, they are also included as a match with the English sentence. This step is shown in Figure 2, where the first sentence is the source English sentence and the corresponding French sentences are the ones with the same length, or length greater than or less than by a value of 2, as compared to the length of the source English sentence.

2.3 Dictionary creation using Anymalign Algorithm

WMT2016 provided us with a large English-French parallel corpus. We executed the Anymalign algorithm on this corpus to find out the word alignments. The alignments with a matching probability of more than or equal to 0.75 were kept as higher probability results in good translation. The rest of the alignments were discarded. This data served as our dictionary. The snapshot of the dictionary containing the source English words in the left column and the target French words in the right column is shown in Figure 3.

```
and et
-    -
Reply    Reply
English English
Equipement   Equipement
NauticNewsletter     NauticNewsletter
Congressional cup    Congressional Cup
Canada.ca    Canada.ca
welcome Accueil
&amp;    &
LBYC     LBYC
NauticNews   NauticNews
glossary     Glossaire
vessels bateaux
:    :
```

Figure 3: Dictionary creation using Anymalign algorithm

2.4 Sentence matching using dictionary

For each of the words in the source English sentence, its corresponding translations are found out using the dictionary produced in the previous step. The words found were then matched with words in the various French sentences that we obtained using the concept provided by Gale and Church. The French sentences, with matched words equal to the length of the source English sentences or less by a factor of 2, were kept and the rest were discarded. This means that for an English sentence of 10 words, French translation for each of the English words were found out using the dictionary produced in the previous step.

If a French sentences with all the 10 words matching to the translated words was found, it was kept. Also, if there was a French sentence con-

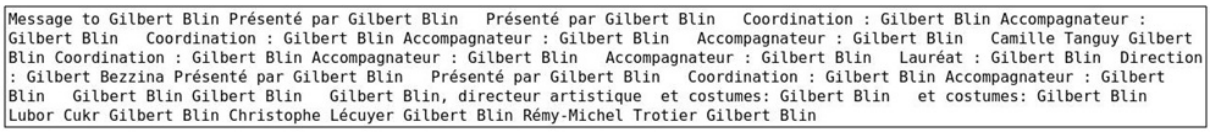

Figure 4: Sentence matching with the derived dictionary

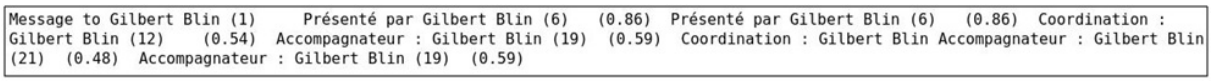

Figure 5: Exact Text Translation finding with Cosine Similarity

taining 10 words, but only 8 words words were matching to the translated words, it was also kept. French sentences with less number of matchings were discarded. This process is shown in Figure 4.

2.5 Exact Text Translation finding with Cosine Similarity

Out of the French sentences extracted in the previous step, Cosine Similarity is found out with respect to the source English sentence. The French sentence with the highest Cosine Similarity score is selected as the exact translation of the source English sentence. This process is shown in Figure 5.

2.5.1 Cosine Similarity

Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1]. The formula used in our approach is as follows.

$$Similarity = cos(\Theta) = \frac{A.B}{\|A\| \, \|B\|} = \frac{\sum_{i=1}^{n} AiBi}{\sqrt{\sum_{i=1}^{n} Ai^2} \sqrt{\sum_{i=1}^{n} Bi^2}} \quad (1)$$

Where A and B are the source English sentence and the one of the target French sentences, respectively.

2.6 URL matching

The URL of the source English sentence is then matched the URL of the extracted French sentence with reference to the ID that was given in the first step. We can see from Figure 5 that the French sentences when compared to the source English sentence "*Message to Gilbert Blin*" have cosine similarity scores appended to it. From the above figure we see that "*Presnt par Gilbert Blin*", has the highest cosine similarity score. So, this can be treated as the exact translation of the source English sentence. We also see that an ID "*(1)*" is appended to the English sentence and an ID "*(6)*" is appended to the French sentence. From Figure 1, we can find out that, since the English sentence ID is "*(1)*", it belongs to the webpage "*http://academiedesprez.org/mailgb.php*" and since the ID of the French sentence is "*(6)*", it belongs to the webpage "*http://academiedesprez.org/eng/musicales5eng.htm*". Thus, we can mark it as the exact alignment.

3 Evaluation

WMT 2016 provided us with a baseline system that finds 119979 extracted pairs after enforcing the 1-1 rule. Our proposed system when executed on the test data, found out 48 extracted pairs of URLs after enforcing the 1-1 rule. This gave our proposed system a percent recall value of 1.998335.

Systems	Extracted pairs
WMT2016 Baseline	119,979
Proposed System	48
Percent Recall	1.998335

Table 1: Evaluation of proposed system with baseline system provided by WMT2016.

4 Conclusion

The paper presents a hybrid approach to bilingual document alignment to the shared task proposed by WMT2016. We have developed an approach that uses the concept given by Gale and Church with respect to length of source-translated sentences, translation of words using a dictionary created by Anymalign and the concept of Cosine Similarity. Our approach was able to extract 48 pairs of URLs with a percent recall of 1.998335.

References

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, pages 169–176, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kenneth Ward Church. 1993. Char_align: A program for aligning parallel texts at the character level. In *In Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 1–8.

I. Dagan, K. Church, and W. Gale, 1999. *Natural Language Processing Using Very Large Corpora*, chapter Robust Bilingual Word Alignment for Machine Aided Translation, pages 209–224. Springer Netherlands, Dordrecht.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, pages 177–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jagadeesh Jagarlamudi, Hal Daumé, III, and Raghavendra Udupa. 2011. From bilingual dictionaries to interlingual document representations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 147–152, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martin Kay and Martin Röscheisen. 1993. Text-translation alignment. *Comput. Linguist.*, 19(1):121–142, March.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, pages 17–22, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adrien Lardilleux, Franois Yvon, and Yves Lepage. 2012. Hierarchical sub-sentential alignment with Anymalign. pages 279–286, Trento, Italy.

Yuji Matsumoto, Hiroyuki Ishimoto, and Takehito Utsuro. 1993. Structural matching of parallel texts. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, pages 23–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, and Sudip Kumar Naskar. 2014. Automatic building and using parallel resources for smt from comparable corpora. *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra)@ EACL*, pages 48–57.

Alexandre Patry and Philippe Langlais, 2005. *Automatic Identification of Parallel Documents With Light or Without Linguistic Resources*, pages 354–365. Springer Berlin Heidelberg, Berlin, Heidelberg.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *in Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 403–411, Stroudsburg, PA, USA. Association for Computational Linguistics.

Liling Tan and Santanu Pal. 2014. Manawi: Using multi-word expressions and named entities to improve machine translation. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 201–206.

Thuy Vu, Ai Ti Aw, and Min Zhang. 2009. Feature-based method for document alignment in comparable news corpora. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 843–851, Stroudsburg, PA, USA. Association for Computational Linguistics.

Susan Warwick, Jan Hajič, and Graham Russell. 1990. Searching on tagged corpora: Linguistically motivated concordance analysis.

Dekai Wu. 1994. Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 80–87, Stroudsburg, PA, USA. Association for Computational Linguistics.