MultiNews: A Web collection of an Aligned Multimodal and Multilingual Corpus

Haithem Afli, Pintu Lohar and Andy Way
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
{FirstName.LastName}@adaptcentre.ie

Abstract

Integrating Natural Language Processing (NLP) and computer vision is a promising effort. However, the applicability of these methods directly depends on the availability of a specific multimodal data that includes images and texts. In this paper, we present a collection of a Multimodal corpus of comparable document and their images in 9 languages from the web news articles of Euronews website.[1] This corpus has found widespread use in the NLP community in Multilingual and multimodal tasks. Here, we focus on its acquisition of the images and text data and their multilingual alignment.

1 Introduction

Although many Natural Language Processing (NLP) applications can be developed by using existing corpora, there are many areas where NLP could be useful if there was a suitable corpus available. For example, Multimodal Machine Translation and Crosslingual Image Description Generation tasks [2] are becoming interested in developing methods that can use not only the texts but also their relations with images. Such information can neither be obtained from standard computer vision data sets such as the COREL collection [3]nor from NLP collections such as Europarl [4] (Koehn, 2005). Similarly, although the image near a text article on a website may provide cues about finding more monolingual and multilingual comparable documents and information on the same topic of the article. We therefore set out to collect a corpus of images aligned with simple full-sentence texts in different languages.

This paper describes our experiences with acquising and aligning multimodal data. Although we did not set out to run a scientific experiment comparing different strategies of how to collect images and texts, our experience points towards certain recommendations for how to collect data for computer vision and NLP domains from news websites such Euronews.

2 Building Multimodal and Multilingual Corpus

The construction of a multilingual corpus for the use in a NLP application typically takes five steps:

(i) obtain the raw data (e.g., web pages)

(ii) align the articles (document alignment)

(iii) extract the texts

(iv) prepare the corpus for NLP applications (normalisation, tokenisation)

(v) map sentences/phrases in one language sentences in the other language (parallel data extraction) In the following, we will describe in detail the acquisition of the Euronews corpus from the website of Euronews.

In this work, data is extracted from the available news (image and text modalities) on the *Euronews* website.[5] Figure 1 shows an example of multimodal comparable data coming from the *Euronews* website. An image source of a political news item and its text version – both in English

[1] euronews.com

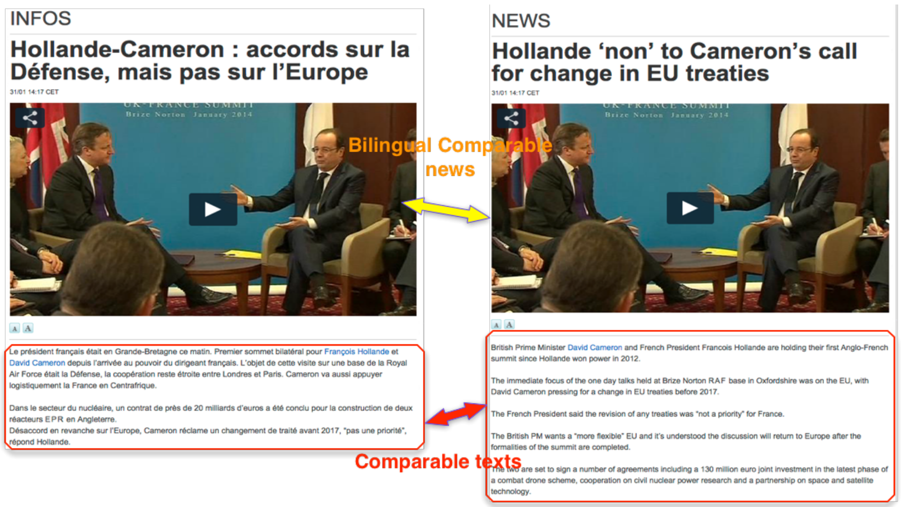
[2] statmt.org/wmt16/multimodal-task.html

[3] The COREL Database for COntent based image REtrievaL https://sites.google.com/site/dctresearch/Home/content-based-image-retrieval

[4] www.statmt.org/europarl/

[5] http://www.euronews.com

Proceedings of the Workshop on Curation and Applications of Parallel and Comparable Corpora, pages 11–15,
Taipei, Taiwan, November 27–December 1, 2017. ©2017 AFNLP



INFOS

Hollande-Cameron : accords sur la Défense, mais pas sur l’Europe

31/01 14:17 CET

Le président français était en Grande-Bretagne ce matin. Premier sommet bilatéral pour François Hollande et David Cameron depuis l'arrivée au pouvoir du dirigeant français. L'objet de cette visite sur une base de la Royal Air Force était la Défense, la coopération reste étroite entre Londres et Paris. Cameron va aussi appuyer logistiquement la France en Centrafrique.

Dans le secteur du nucléaire, un contrat de près de 20 milliards d'euros a été conclu pour la construction de deux réacteurs EPR en Angleterre.
Désaccord en revanche sur l'Europe, Cameron réclame un changement de traité avant 2017, "pas une priorité", répond Hollande.

NEWS

Hollande ‘non’ to Cameron’s call for change in EU treaties

31/01 14:17 CET

British Prime Minister David Cameron and French President Francois Hollande are holding their first Anglo-French summit since Hollande won power in 2012.

The immediate focus of the one day talks held at Brize Norton RAF base in Oxfordshire was on the EU, with David Cameron pressing for a change in EU treaties before 2017.

The French President said the revision of any treaties was "not a priority" for France.

The British PM wants a "more flexible" EU and it's understood the discussion will return to Europe after the formalities of the summit are completed.

The two are set to sign a number of agreements including a 130 million euro joint investment in the latest phase of a combat drone scheme, cooperation on civil nuclear power research and a partnership on space and satellite technology.

Figure 1: Example of comparable documents from the *Euronews* Web site.

– are available along with the equivalent news in French (image and text modalities). These documents can be used to extract comparable documents and parallel data.

Euronews web site clusters news into several categories including languages and sub-domains (*e.g.* Sport, Politics, etc.). Table 2 shows the statistics of our *MMEuronews* corpus created from news article data from 2013, 2014 and 2105 in 9 languages including: fr(French), ar(Arabic), en(English), de(German), es(Spanish), it(Italian), tr(Turkish), ua(Ukrainian), and pt(Portuguese).

3 Aligning Comparable documents

3.1 Basic Idea

We propose an extension of the method described in (Sennrich and Volk, 2010) to align our corpus. The basic system architecture is described in Figure 2. We begin by removing the documents that have very little contents in order to reduce the total number of all possible comparisons. Such documents are very rarely considered as candidates for being comparable document because they consist of only few sentences or words and it is observed that in the reference for training data provided, these kind of documents are not included in the reference set. Subsequently, we introduced three methods as follows: (i) sentence-level scoring, (ii) word-level scoring, and (iii) named entity (NE)-based scoring.

Finally we added these three scores to select the 1-best target document which has the highest value.

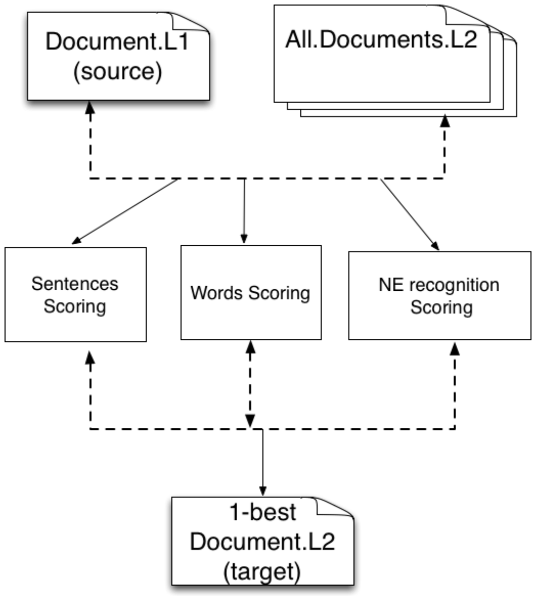


Figure 2: Architecture of comparable alignment system alignment

Language	en	fr	ar	de	es	it	pt	tr	ua	Total
# Articles	40421	39663	36836	37293	37218	36970	36854	37291	37021	339 567

Table 1: Size of the Euronews transcribed English audio corpus and English-French texts.

3.2 Sentence based scoring

Since there are a large number of source and target documents especially in the domains with with a large amount of documents, we have to restrict the comparison process only to the source-target document pairs that have close sentence-length ratio. Otherwise they are very less likely to be comparable documents. It is necessary since comparing each source with each target document would result in an undesirably large number of comparisons($m * n$, with m and n being the total number of source and target documents, respectively in a specific domain) and therefore very long time for the whole computation even for a single domain. Let us assume that S_s and S_t are number of sentences in source and target document respectively. Assuming this we follow very simple formula to calculate source-target sentence-length ratio(S_{LR}) as follows:

$$S_{LR} = \frac{Min(S_s, S_t)}{Max(S_s, S_t)} \quad (1)$$

We construct this equation in order to confine the value between 0 and 1 which implies that if either of source and target document has no sentences, S_{LR} will be 0 and 1 if they have same number of sentences. Therefore, a value of 1 or even very close to it has positive indication towards being comparable but this is not the only requirement as there are many documents with same or nearly same number of sentences. Due to this reason, we consider word and NE-based scorings in sections 3.4 and 3.5 respectively.

3.3 Word based scoring

The reason behind using this method is very similar to the method discussed in Section 3.2 except that it is used at word level. Let us assume that W_s and W_t are number of words in source and target documents respectively. Hence our equation for calculating source-target word-length ratio(W_{LR}) becomes:

$$W_{LR} = \frac{Min(W_s, W_t)}{Max(W_s, W_t)} \quad (2)$$

3.4 NE-based scoring

After a linguistic study on the comparable documents, we found that looking for NEs present in both source and target documents can be a good way to select the 1-best target document. We extracted NEs from all the documents to be compared and calculate the percentage of source NE matches(P_{SNM}) with target NEs.

However, in many cases a source and a target documents can have huge difference in number of NEs. For example, if a source document has 5 and a target document has 50 NEs respectively and all of the source NEs match with target NEs, it is probably a bad idea to simply calculate P_{SNM} and add to the sentence-based and to the word-based scores. Due to this reason we consider the source-target-NE-length ratio (NE_{LR}) and multiplied it with P_{SNM} . Hence the weight of P_{SNM} is decreased from 100% to 10% which is a result from depending upon N_{LR}. Henceforth, the NE-based score(NE_{SC}) is described as:

$$NE_{SC} = P_{SNM} * N_{LR} \quad (3)$$

3.5 Combining all scores

We propose to re-rank our possible alignments based on adding sentence, word and NE-based scores and call this as alignment-score (A_{SC})

$$A_{SC} = S_{LR} + W_{LR} + NE_{SC} \quad (4)$$

Using equation 4 we calculate scores for each document pairs in comparison and retain the 1-best pair that has the maximum value.

4 Results

The results are given in Table 2. Each row in the table contains three numerical values that represent (from left to right) the total numbers of source-language, target-language and aligned document pairs, respectively. As we can see, we are successfully aligning images and texts in 8 pair of languages. We produced a total of more than 288k of bilingual aligned multimodal documents. Our corpus, alignment model and code will be made

Documents	# Source	# Target	# Aligned
En-Ar	40421	36836	35761
En-De	40421	37293	36114
En-Es	40421	37218	36178
En-Fr	40421	37293	36762
En-It	40421	36970	36003
En-Pt	40421	36854	35863
En-Tr	40421	37291	35901
En-Ua	40421	37021	35922
Total			288 504

Table 2: Results of bilingual aligned image-text MMEuronews data used in our experiments.

publicly for the computer vision and NLP community.

5 Related Work on Document Alignment

In the "Big Data" world that we now live in, it is widely believed that *there is no better data than more data* (e.g. Mayer-Schönberger and Cukier (2013)). In line with this idea, many works use the Web as resource for building corpus for document alignment and parallel text extraction tasks. However, the extensive literature related to the problem of exploiting comparable corpora takes a somewhat different perspective than we do in this paper.

Typically, comparable corpora do not have any information regarding document pair similarity. They are made of many documents in one language which do not have any corresponding translated document in the other language. Furthermore, when the documents are paired, they are not literal translations one of each other. Thus, extracting parallel data from such corpora requires special algorithms.

Many works use the Web as a comparable corpus. An adaptive approach, proposed by Zhao and Vogel (2002), aims at mining parallel sentences from a bilingual comparable news collection collected from the Web. A maximum likelihood criterion was used by combining sentence-length models with lexicon-based models. The translation lexicon is iteratively updated using the mined parallel data to obtain better vocabulary coverage and translation probability estimation. Resnik and Smith (2003) propose a web-mining-based system called STRAND and show that their approach is able to find large numbers of similar document pairs. In (Yang and Li, 2003), an alignment method is presented at different levels (title, word and character) based on dynamic programming (DP). The goal is to identify one-to-one title pairs in an English–Chinese corpus collected from the Web. They apply the longest common subsequence to find the most reliable Chinese translation of an English word.

One of the main methods relies on cross-lingual information retrieval (CLIR), with different techniques for transferring the request into the target language (using a bilingual dictionary or a full SMT system). Utiyama and Isahara (2003) use CLIR techniques and DP to extract sentences from an English–Japanese comparable corpus. They identify similar article pairs, and having considered them as parallel texts, then align sentences using a sentence-pair similarity score and use DP to find the least-cost alignment over the document pair. (Munteanu and Marcu, 2005) use a bilingual lexicon to translate some of the words of the source sentence. These translations are then used to query the database to find matching translations using IR techniques.

There have been only a few studies trying to investigate the formal quantification of how similar two comparable documents are. Li and Gaussier (2010) presented one of the first works on developing a comparability measure based on the expectation of finding translation word pairs in the corpus. Our approach follows this line of work based on a method developed by Sennrich and Volk (2010).

6 Conclusion

Despite the fact that many researchers have investigated the use of comparable corpora to generate initial training data for NLP, we still have a lack of corpus in different modalities.

In this paper, we seek to build a corpus that combine aligned images and texts in different languages. We use Euronews website as source of our crawled raw data. We propose a new techniques to align bilingual documents. Our method is based on Matched source-to-target sentence/words and Named Entity scoring.

Given this promising result, in future work we would like to add more language pairs and data to our corpus. In addition, we wish to investigate its utility to improve the extraction of parallel data from multilingual comparable corpora. We plan, also, to develop a new model for building embeddings that are both multilingual and multimodal.

Finally, we hope that the availability of this kind of resources (corpora, tools) continues to make computer vision and NLP an exciting and productive fields.

Acknowledgments

This research is supported by Science Foundation Ireland through ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University.

References

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of MT Summit*.

Li, B. and Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010),*, pages 644–652.

Mayer-Schönberger, V. and Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. In *London: Hodder & Stoughton*.

Munteanu, D. S. and Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

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

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

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 72–79.

Yang, C. C. and Li, K. W. (2003). Automatic construction of english/chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, 54:730–742.

Zhao, B. and Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, Washington, DC, USA. IEEE Computer Society.