The ILSP/ARC submission to the WMT 2016 Bilingual Document Alignment Shared Task

Vassilis Papavassiliou **Prokopis Prokopidis** **Stelios Piperidis**
Institute for Language and Speech Processing
Athena Research and Innovation Center
Athens, Greece
{vpapa, prokopis, spip}@ilsp.gr

Abstract

This paper describes ILSP-ARC-pv42, the Institute for Language and Speech Processing/Athena Research and Innovation Center submission for the WMT 2016 Bilingual Document Alignment shared task. We describe several document and collection-aware features that our system explored in the context of the task. On the test dataset, our submission achieved a recall of 84.93%, even though it does not make use of any language-specific resources like bilingual lexica or MT output. Instead, our system is based on shallow features (including links to documents in the same webdomain, URLs, digits, image filenames and HTML structure) that can be easily extracted from web documents. We also present examples to show that when de-duplication issues in the test dataset are properly addressed, our system reaches a significantly higher recall of 92.5%.

1 Introduction

There is a growing literature on using web-acquired data for constructing various types of language resources, including monolingual and parallel corpora. As shown in, among others, Pecina et al. (2014) and Rubino et al. (2015), such resources can be exploited in training generic or domain-specific machine translation systems. Nevertheless, compared to the acquisition of monolingual data from the web, construction of parallel resources is more challenging. Even though there are many multilingual websites with pairs of documents that are translations of each other, detection of such sites and identification of the document pairs is far from straightforward. Resnik and Smith (2003) presented the STRAND system, in which they used a search engine to search for multilingual websites and examined the similarity of the HTML structures of the fetched webpages in order to identify pairs of potentially parallel pages. Esplà-Gomis and Forcada (2010) developed Bitextor, a system that combines language identification with shallow features. Barbosa et al. (2012) crawl the web and examine the HTML DOM tree of visited webpages with the purpose of detecting multilingual websites based on the collation of links that are very likely to point to in-site pages in different languages. Smith et al. (2013) used an extension of the STRAND algorithm to perform large-scale experiments of mining parallel documents from the Common Crawl[1] dataset.

This paper describes ILSP-ARC-pv42, the Institute for Language and Speech Processing/Athena Research Center submission for the WMT 2016 Bilingual Document Alignment shared task. The task consisted in identifying pairs of English and French documents from collections of documents corresponding to crawls of 203 webdomains.

2 System architecture

In this section, we describe the main processing steps in ILSP-ARC-pv42. Our system is based on the document alignment module of the ILSP Focused Crawler (Papavassiliou et al., 2013), an open-source tool[2] that integrates all necessary software[3] for the creation of high-precision parallel resources from the web in a language-independent fashion.

[1] http://commoncrawl.org/

[2] http://nlp.ilsp.gr/redmine/ilsp-fc/

[3] Including modules for metadata extraction, language identification, boilerplate removal, document clean-up, text classification and sentence alignment

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 733–739,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

2.1 Pre-processing shared task files

We pre-processed crawled data provided by the organizers as one file per webdomain in the `.lett` format adapted from Bitextor. This is a plain text format with one line per web document. Each line consists of 6 tab-separated values that include the (automatically detected) language ID ({en, fr}); the mime type (always text/html); the encoding (always charset=utf-8); the URL; the HTML content in Base64 encoding; and the text in Base64 encoding.

For each webdomain, we created a directory where we exported the contents of the 5th field of each entry in a *ll-file_id.html* file, where *ll* is the two letter language id ({en, fr}) provided in the lett files and *file_id* is an integer unique for each file of a webdomain. Using the URL information, we also store file-to-URL mappings in a separate file.

Apart from training and test data in this format, the organizers also identified spans of FR text for which they produced EN translations using a machine translation system. In an attempt to recreate real-life conditions where, at least for our team and for many language pairs, access to reliable MT output is not available, we did not use this information or any other type of language- or language-pair-dependent information in our system.

2.2 Boilerplate detection and exporting

Apart from its textual content, a typical webpage also contains boilerplate, i.e. "noisy" elements like navigation headers, advertisements, disclaimers etc., which are of only limited or no use for the production of good-quality language resources. We used a modified version of Boilerpipe[4] (Kohlschtter et al, 2010) to identify boilerplate in the `.html` files. Besides boileplate detection, we also identified structural information like *title*, *heading* and *list item* from each webpage. At this stage, text was also segmented into paragraphs on the basis of specific HTML tags like `<p>`, `</br>`, `<li>` etc.

For each `.html` file, we generated an `.xml` file where a `<body>` element contained the content of the document segmented into paragraphs. Apart from normalized text, each paragraph element was enriched with attributes providing more information about the process outcome. Specifically, paragraphs may contain the following attributes: i) *crawlinfo* with possible values *boilerplate*, meaning that the paragraph has been considered boilerplate; and ii) *type* with possible values: *title, heading* and *listitem*.

2.3 Document pair detection

Following exporting, a document pair detector, which constitutes the core module of our system, applies a set of complementary methods based on the content of the `.html` and the `.xml` files in order to identify translation pairs. The module does not exploit any language resources (e.g. lexica or output of MT engines). Instead it is based on shallow features including links to documents in the same webdomain, URLs, digits, image filenames and HTML structure.

We trivially avoid pairing files that are in the same language. We then examine all links in the `.html` files and we extract those that contain the `hreflang` attribute. Since "hreflang specifies the language and optional geographic restrictions for a document"[5], we use this strong indicator to pair documents, which we subsequently exclude from examination by other downstream methods[6]. We also examine links that match a set of patterns for the identification of translation links (e.g. link elements with the attribute `lang`) and we exploit them in the same way.

Next, we focus on URLs that include language indicators and examine if there are pairs of URLs that match pairs of specific patterns such as */lang1/* and */lang2/*, *_lang1* and *_lang2*, *=lang1* and *=lang2*, where *lang1* and *lang2* are alternative representations of the targeted languages (e.g. en, eng, english, fr, fra, french, francais, etc. in the context of this shared task). Some additional patterns are lang=*i*, langid=*i* and lingua=*i*, where $i \in \{0, .., 5\}$.

It is worth mentioning that in the past we have complemented the use of the above indicators with examination of features like document length in terms of tokens/paragraphs, in order to decide on document pairness. This was in accordance with our main interest in using the pair detector for the generation of high-quality resources that can be used in improving MT systems. However, in the context of this recall-evaluated shared task, the

[4] `http://code.google.com/p/boilerpipe/`

[5] `https://en.wikipedia.org/wiki/Hreflang`

[6] We use this approach for all methods: documents that have been paired by one method are excluded from further examination.

system bases its decision on these indicators without any further checks.

Then, each `.xml` file is parsed and the following features are extracted: i) the document *language*[7]; ii) the *depth* of the original source page, (e.g. for `http://domain.org/d1/d2/d3/page.html`, depth is 4); iii) the *number of paragraphs*; iv) the *length* (in terms of tokens) of the main content, i.e. non-boilerplate text; v) the sequence of digits in the main content; and vi) the *fingerprint* of the main content, which is a sequence of integers that represent the structural information of the page, with boilerplate content ignored. For instance, in a fingerprint of $[-2, 28, 145, -4, 9, -3, 48, 740]$ for a document of 6 paragraphs, negative numbers -2, -3 and -4 denote that the *type* attributes of the 1st, 3rd and 4th `<p>` elements have *title*, *heading* and *listitem* values, respectively; and positive integers are the lengths of the 6 paragraphs in characters.

At this stage, webpages with a depth difference > 1 are not examined as candidate translations of each other, on the assumption that it is unlikely that translations can be found at very distant levels of the web site tree.

We next extract the filenames of the images from the HTML source and each document is represented as a list of images[8]. Our assumption at this stage is that two documents that contain the same or a similar set of images are good candidates for pairing. Since it is very likely that some images appear in many webpages, we count the occurrence frequency of each image and we discard "common", i.e. relatively frequent, images (e.g. social media icons, logos etc.) from these lists.

In order to classify images into "critical" or "common" (see Figure 1) we need to calculate a threshold. In principle, one should expect that low/high frequencies correspond to "critical"/"common" images. We employ a non-parametric approach for estimating the probability density function (Alpaydin, 2010) of the image frequencies using the following formula:

$$\hat{p}(x) = \frac{1}{Mh} \sum_{t=1}^{M} K\left(\frac{x-x^t}{h}\right)$$

where the random variable x defines the positions (i.e. images frequencies) at which the $\hat{p}(x)$ will be estimated, M is the amount of images, x^t denotes the values of data samples in the region of width h around the variable x, and $K(\cdot)$ is the normal kernel that defines the influence of values x^t in the estimation of $\hat{p}(x)$. The optimal value for h, i.e. the optimal bandwidth of the kernel smoothing window, was calculated as described in Bowman and Azzalini (1997).

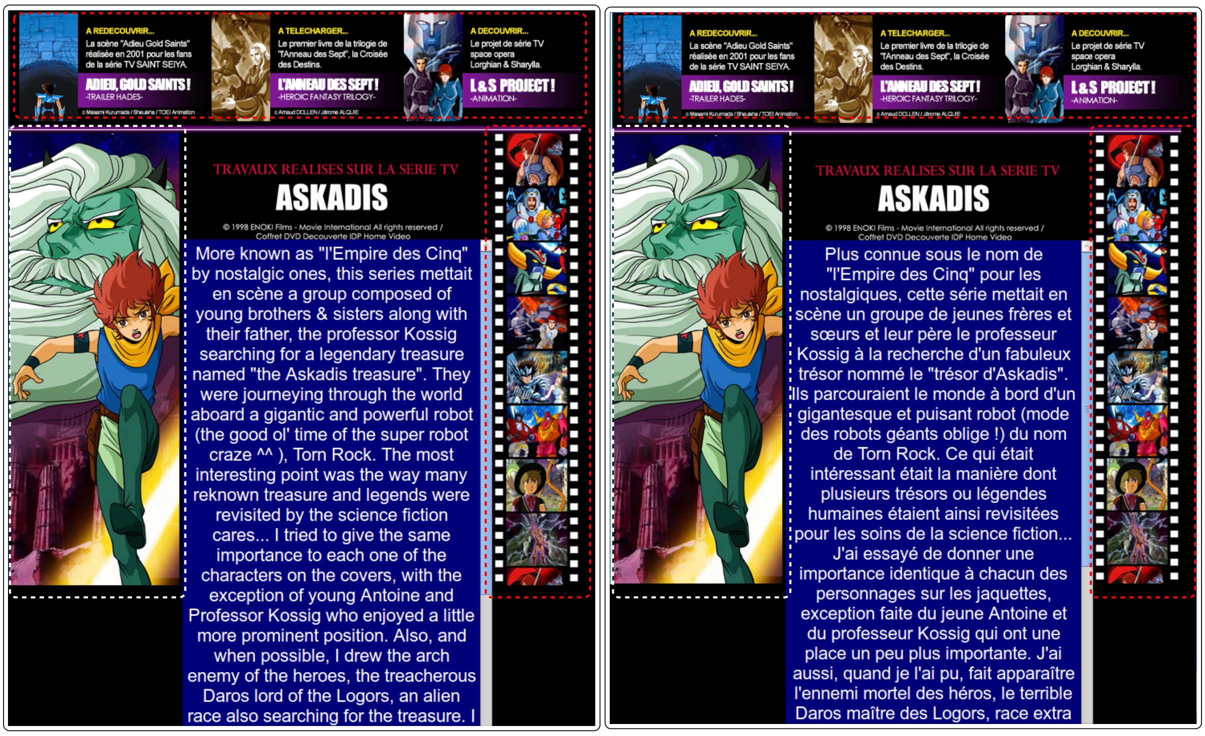
Figure 2 serves as an illustration of the normalized histogram of image frequencies in an example webdomain (that was not part of the shared task datasets) and the estimated probability density function. One can identify a main lobe in the low values, around which ”critical” images are clustered. Thus, the threshold is chosen to be equal to the minimum just after this lobe. The underlying assumption is that if a webpage in *l1* contains image(s), then the webpage with its translation in *l2* will contain a similar set of images. In case this assumption is not valid for a multilingual webdomain (i.e. if there are only images that appear in all pages, e.g. template icons), then all images will wrongly be assumed to be "critical". To eliminate this problem, we also discard as "common" all images that appear in more than 10% of the total `.html` files of each webdomain.

Following this step, each document is examined against all others on the basis of: a) the Jaccardian similarity coefficient of their image lists b) the reciprocal of edit distance of the sequences of digits in their main content c) the ratio of their number of paragraphs and d) the ratio of the number of tokens in non-boilerplate text. Two documents are considered parallel if (c), (d) and either or both of (a) and (b) are above predefined thresholds.

Additional document pairs are detected by examining structure similarity. Since the `.xml` files contain information about both (non-boilerplate) content *and* structure (i.e. titles, headings, list items), we use this representation instead of examining the similarity on the actual HTML source. A 3-dimensional feature vector is constructed for each candidate pair of parallel documents. The first element in this vector is the ratio of their fingerprint lengths, the second is the ratio of their paragraph size, and the third is the ratio of the edit distance of the fingerprints of the two documents to the maximum fingerprint length. Classification of a pair as parallel is performed using a

[7] In the case of the shared task, we replace the output of a language detection module with the language id provided by the organizers.

[8] Henceforth, we use the term "image" to mean "image filenames". We do not make use of any image features other than their filenames.



A REDECOUVRIR...
La scène "Adieu Gold Saints" réalisée en 2001 pour les fans de la série TV SAINT SEIYA.
ADIEU, GOLD SAINTS !
-TRAILER HADES-
© Masami Kurumada / Shueisha / TOEI Animation

A TELECHARGER...
Le premier livre de la trilogie de "l'Anneau des Sept", la Croisée des Destins.
L'ANNEAU DES SEPT !
-HEROIC FANTASY TRILOGY-
© Arnaud DOLLEN / Jérome ALQUIE

A DECOUVRIR...
Le projet de série TV space opera Lorghian & Sharylla.
L&S PROJECT !
-ANIMATION-

TRAVAUX REALISES SUR LA SERIE TV
ASKADIS
© 1998 ENOKI Films - Movie International All rights reserved / Coffret DVD Decouverte IDP Home Video

More known as "l'Empire des Cinq" by nostalgic ones, this series mettait en scène a group composed of young brothers & sisters along with their father, the professor Kossig searching for a legendary treasure named "the Askadis treasure". They were journeying through the world aboard a gigantic and powerful robot (the good ol' time of the super robot craze ^^), Torn Rock. The most interesting point was the way many reknown treasure and legends were revisited by the science fiction cares... I tried to give the same importance to each one of the characters on the covers, with the exception of young Antoine and Professor Kossig who enjoyed a little more prominent position. Also, and when possible, I drew the arch enemy of the heroes, the treacherous Daros lord of the Logors, an alien race also searching for the treasure. I

A REDECOUVRIR...
La scène "Adieu Gold Saints" réalisée en 2001 pour les fans de la série TV SAINT SEIYA.
ADIEU, GOLD SAINTS !
-TRAILER HADES-
© Masami Kurumada / Shueisha / TOEI Animation

A TELECHARGER...
Le premier livre de la trilogie de "l'Anneau des Sept", la Croisée des Destins.
L'ANNEAU DES SEPT !
-HEROIC FANTASY TRILOGY-
© Arnaud DOLLEN / Jérome ALQUIE

A DECOUVRIR...
Le projet de série TV space opera Lorghian & Sharylla.
L&S PROJECT !
-ANIMATION-

TRAVAUX REALISES SUR LA SERIE TV
ASKADIS
© 1998 ENOKI Films - Movie International All rights reserved / Coffret DVD Decouverte IDP Home Video

Plus connue sous le nom de "l'Empire des Cinq" pour les nostalgiques, cette série mettait en scène un groupe de jeunes frères et sœurs et leur père le professeur Kossig à la recherche d'un fabuleux trésor nommé le "trésor d'Askadis". Ils parcouraient le monde à bord d'un gigantesque et puisant robot (mode des robots géants oblige !) du nom de Torn Rock. Ce qui était intéressant était la manière dont plusieurs trésors ou légendes humaines étaient ainsi revisitées pour les soins de la science fiction... J'ai essayé de donner une importance identique à chacun des personnages sur les jaquettes, exception faite du jeune Antoine et du professeur Kossig qui ont une place un peu plus importante. J'ai aussi, quand je l'ai pu, fait apparaître l'ennemi mortel des héros, le terrible Daros maître des Logors, race extra

Figure 1: Critical (white) and common (red) images in two documents from the `www.jerome-alquie.com.lett` webdomain.

soft-margin polynomial Support Vector Machine trained with the positive and negative examples collected in the context of previous experiments (Pecina et al., 2012).

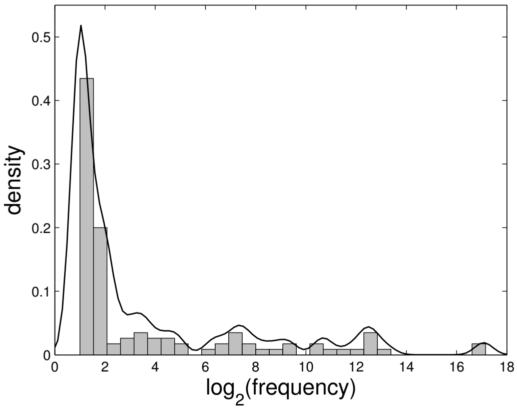


Figure 2: The normalized histogram and the estimated pdf of image frequencies in an example webdomain

As a final step, we mapped each *ll-file_id.html* to its URL and we produced a final set of 291,749 proposed pairs for all webdomains of the test data.

3 Evaluation Results

Before submitting our proposed pairs on the shared task test data, we also evaluated our system on the training data. The latter consisted of a set of 1,624 EN-FR pairs extracted by the organizers from 49 webdomains. The number of pairs per webdomain in the training set varied between 4 and over 230. The simple baseline provided by the organizers is based on URL matching. The baseline implementation iterates through all URLs and strips language identifiers such as */english/* from URLs. It then produces pairs of URLs that have the same stripped representation. Overall, the baseline proposes 143,851 candidate pairs, which are reduced to 119,979 pairs after enforcing the 1-1 rule, which requires that each source URL may be matched with at most one target url and vice-versa. Should a URL occur repeatedly, later occurrences are ignored. The baseline identifies 1,103 true positives, thus reaching a recall of 67.92%. Our system proposed 160,727 EN-FR pairs from which 1,460 are included in the EN-FR training set pairs, corresponding to a 89.90% recall on the training dataset.

Following our submission of predicted pairs on the shared task test data, the organizers evaluated it against a set of 2,402 EN-FR pairs from the 203 webdomains comprising the test data. The number of pairs per webdomain in the test data varies between 1 and 357, while the number of EN and FR webpages of each webdomain varies between 5 and circa 99K.

Our system proposed 291,749 pairs that were reduced to 287,860 after enforcing the 1:1 rule. These additional pairs were created because, for certain domains, EN or FR webpages contained translation links pointing to multiple webpages identified by the organizers as FR or EN documents, respectively. Our system identified 2,040 EN-FR pairs out of the 2,402 provided test pairs at a 84.93% recall and was ranked 9th among the 21 submitted systems by the 13 participant groups.

We counted the number of true positive pairs identified via each method, in order to examine each method's contribution. The top contributing method with 987 (48.38%) of the correctly detected pairs was the one exploiting URL patterns. Methods based on the existence of common images and/or similar digit sequences contributed 791 pairs (38.77%) while the in-webdomain links and HTML structure generated 180 (8.82%) and 82 (4.02%) pairs, respectively.

We also examined manually all document pairs missed in our submission in order to gather useful insights that could help us improve our system. A first conclusion is that a major issue in evaluating bilingual document alignment in terms of recall concerns (near) duplicates. We observed that we were scored as missing 182 pairs because the EN and/or FR documents participating in each of these pairs were aligned by our system with documents that contained the same content but originated from different URLs. For example, 50 and 103 test pairs from the `www.taize.fr` (see Figure 3) and the `www.lalettrediplomatique.fr` webdomains (where extra attribute-value strings in URLs like `choixlang=1`, `&bouton=1`, `&bouton=2`, etc. do not "influence" the content of the FR webpages) were considered fails due to this issue in the test data. Additional examples of perfectly valid pairs for extracting valuable content for downstream MT applications, which a) have been proposed by our system b) are equivalent to test pairs but c) have not been scored as true positives, are presented in Table 1. In particular, the `www.lagardere.com` and the `www.zigiz.com` webdomains (rows 7 and 8) contribute 11 and 6 missed pairs, respectively. If we consider all these pairs as valid for extracting data in order to train MT systems, our system reaches a recall of 92.5%.

The majority of the remaining $(2402 - 2040 - 182 =)$ 180 of our misses concerned pairs where for a page A, the method based on structure similarity proposed a wrong document pair with page B. For instance, in the `http://www.toucherdubois.ca` webdomain, information (concerning learning scenarios and teaching resources) is presented in a specific format/template leading to errors during the examination of the structure fingerprint. Other misses were due to the length of the documents since it is very identify pairs of very short documents without using any lexical information.

4 Conclusions

In this paper we described the ILSP/ARC submission for the WMT 2016 Bilingual Document Alignment shared task. We provided details on the document and collection-aware features that our system explores. On the test set, our system reached a recall of 84.93% according to the official scoring. In the evaluation section of the paper we presented examples in order to show that the recall of our system is significantly higher once de-duplication issues in the test data are addressed.

Acknowledgments

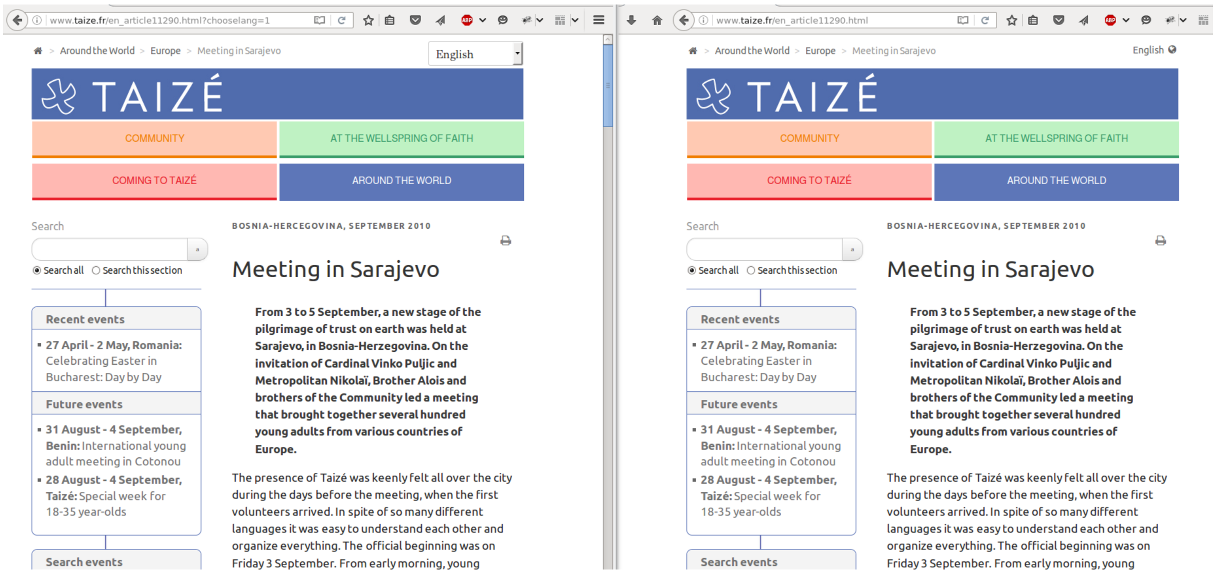
This work was supported by the Abu-MaTran (FP7-People-IAPP, Grant 324414) project and the European Language Resource Coordination effort, CEF Programme. We would like to thank the organizers for preparing the task and answering questions.

References

Ethem Alpaydin. 2010. *Introduction to Machine Learning*. The MIT Press, 2nd edition.

Luciano Barbosa, Vivek Kumar Rangarajan, Sridhar, Mahsa Yarmohammadi, and Srinivas Bangalore. 2012. Harvesting parallel text in multiple languages with limited supervision. In *COLING*, pages 201–214.

Adrian W. Bowman and Adelchi Azzalini. 1997. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. Oxford University Press.


www.taize.fr/en_article11290.html?chooselang=1
Around the World > Europe > Meeting in Sarajevo
English
TAIZÉ
COMMUNITY
AT THE WELLSPRING OF FAITH
COMING TO TAIZÉ
AROUND THE WORLD
Search
○ Search all ○ Search this section

BOSNIA-HERCEGOVINA, SEPTEMBER 2010

Meeting in Sarajevo

From 3 to 5 September, a new stage of the pilgrimage of trust on earth was held at Sarajevo, in Bosnia-Herzegovina. On the invitation of Cardinal Vinko Puljic and Metropolitan Nikolaï, Brother Alois and brothers of the Community led a meeting that brought together several hundred young adults from various countries of Europe.

The presence of Taizé was keenly felt all over the city during the days before the meeting, when the first volunteers arrived. In spite of so many different languages it was easy to understand each other and organize everything. The official beginning was on Friday 3 September. From early morning, young

Recent events

- **27 April - 2 May, Romania:** Celebrating Easter in Bucharest: Day by Day

Future events

- **31 August - 4 September, Benin:** International young adult meeting in Cotonou
- **28 August - 4 September, Taizé:** Special week for 18-35 year-olds

Search events

www.taize.fr/en_article11290.html
Around the World > Europe > Meeting in Sarajevo
English
TAIZÉ
COMMUNITY
AT THE WELLSPRING OF FAITH
COMING TO TAIZÉ
AROUND THE WORLD
Search
○ Search all ○ Search this section

BOSNIA-HERCEGOVINA, SEPTEMBER 2010

Meeting in Sarajevo

From 3 to 5 September, a new stage of the pilgrimage of trust on earth was held at Sarajevo, in Bosnia-Herzegovina. On the invitation of Cardinal Vinko Puljic and Metropolitan Nikolaï, Brother Alois and brothers of the Community led a meeting that brought together several hundred young adults from various countries of Europe.

The presence of Taizé was keenly felt all over the city during the days before the meeting, when the first volunteers arrived. In spite of so many different languages it was easy to understand each other and organize everything. The official beginning was on Friday 3 September. From early morning, young

Recent events

- **27 April - 2 May, Romania:** Celebrating Easter in Bucharest: Day by Day

Future events

- **31 August - 4 September, Benin:** International young adult meeting in Cotonou
- **28 August - 4 September, Taizé:** Special week for 18-35 year-olds

Search events

Figure 3: Duplicate EN webpages from `www.taize.fr`. Proposed pairs including `http://www.taize.fr/en_article11290.html?chooselang=1` (left) – `http://www.taize.fr/fr_article11286.html?chooselang=1` were not considered equivalent to test pairs including `http://www.taize.fr/en_article11290.html` (right) – `http://www.taize.fr/fr_article11286.html` even though the extra `?chooselang=1` attribute-value string in the URLs does not influence the textual content that can be extracted from proposed pairs.

Miquel Esplà-Gomis and Mikel L. Forcada. 2010. Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathemathical Lingustics*, 93:77–86.

Vassilis Papavassiliou, Prokopis Prokopidis, and Gregor Thurmair. 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.

Pavel Pecina, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, and Josef van Genabith. 2012. Domain adaptation of statistical machine translation using web-crawled resources: a case study. In *Proceedings of the 16th Annual Conference of EAMT*, pages 145–152, Trento, Italy.

Pavel Pecina, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, Aleš Tamchyna, Andy Way, and Josef Genabith. 2014. Domain adaptation of statistical machine translation with domain-focused web crawling. *Language Resources and Evaluation*, 49(1):147–193.

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

Raphael Rubino, Tommi Pirinen, Miquel Esplà-Gomis, Nikola Ljubešić, Sergio Ortiz Rojas, Vassilis Papavassiliou, Prokopis Prokopidis, and Antonio Toral. 2015. Abu-matran at wmt 2015 translation task: Morphological segmentation and web crawling. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 184–191, Lisbon, Portugal, September. Association for Computational Linguistics.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st ACL*, pages 1374–1383, Sofia, Bulgaria.

id	type	EN URL	FR URL
1	T	http://www.lucistrust.org/it/service_activities/world_goodwill/world_view_archive/world_view_the_trained_observer	http://www.lucistrust.org/fr/service_activities/world_goodwill/world_view_archive/world_view_the_trained_observer
	S	http://www.lucistrust.org/en/service_activities/world_goodwill/world_view_archive/world_view_the_trained_observer	http://www.lucistrust.org/fr/service_activities/world_goodwill/world_view_archive/world_view_the_trained_observer
2	T	http://www.eufic.org/article/cs/page/BARCHIVE/expid/basics-child-adolescent-nutrition/	http://www.eufic.org/article/fr/page/BARCHIVE/expid/basics-alimentation-enfants-adolescents/
	S	http://www.eufic.org/article/en/page/BARCHIVE/expid/basics-child-adolescent-nutrition/	http://www.eufic.org/article/fr/page/BARCHIVE/expid/basics-alimentation-enfants-adolescents/
3	T	http://www.phytoclick.com/index.html?lang=en&pID=172	http://www.phytoclick.com/conditions-generales-de-vente/index.htm
	S	http://www.phytoclick.com/index.html?lang=en&pID=172&bID=151	http://www.phytoclick.com/index.html?pID=172&bID=151
4	T	http://www.eurovia.org/spip.php?article330	http://www.eurovia.org/spip.php?article329
	S	http://www.eurovia.org/spip.php?article330&lang=fr	http://www.eurovia.org/spip.php?article329&lang=es
5	T	http://www.haro.com/en/cork/all_about_cork/general.php	http://www.haro.com/fr/liege/tout_sur_le_liege/general.php
	S	http://www.haro.com/us/cork/all_about_cork/general.php	http://www.haro.com/fr/liege/tout_sur_le_liege/general.php
6	T	http://www.kinnarps.com/en/International/InteriorSolutions/KinnarpsBenefits/Ergonomics/Light/	http://www.kinnarps.com/fr/ch/Solutions-d-amenagement/Les-avantages-Kinnarps/Ergonomie/Lumiere/
	S	http://www.kinnarps.com/en/uk/InteriorSolutions/Ergonomics/Ergonomics/Light/	http://www.kinnarps.com/fr/ch/Solutions-d-amenagement/Les-avantages-Kinnarps/Ergonomie/Lumiere/
7	T	http://www.lagardere.com/press-room/press-releases/press-releases-363.html&idpress=1268	http://www.lagardere.com/centre-presse/communiques-de-presse/communiques-de-presse-122.html&idpress=3168
	S (11)	http://www.lagardere.com/press-room/press-releases/press-releases-363.html&idpress=1268	http://www.lagardere.com/press-room/press-releases/press-releases-363.html&idpress=3168
8	T	http://www.zigiz.com/en-EN/help/about_zigiz/help_parent_actievoorwaarden.html	http://www.zigiz.com/fr-FR/aide/about_zigiz/help_parent_actievoorwaarden.html
	S (6)	http://www.zigiz.com/en-EN/help/about_zigiz/help_parent_actievoorwaarden.html	http://www.zigiz.com/fr-FR/aide/help_parent_faq/help_allpaymentmethods.html
9	T	http://www.oras.com/en/professional/products/Pages/ProductVariant.aspx?productcode=6527A	http://www.oras.com/be/professional/products/Pages/ProductVariant.aspx?productcode=6527A
	S	http://www.oras.com/en/professional/products/Pages/ProductVariant.aspx?productcode=6527A	http://www.oras.com/fr/professional/products/Pages/ProductVariant.aspx?productcode=6527A
10	T	http://www.ipu.org/hr-e/169/Co121.htm	http://www.ipu.org/hr-f/168/Co121.htm
	S	http://www.ipu.org/hr-e/169/Co121.htm	http://www.ipu.org/hr-f/169/Co121.htm
11	T	http://www.nserc-crsng.gc.ca/Prizes-Prix/Excellence-Excellence/Profiles-Profils_eng.asp?ID=1008	http://www.nserc-crsng.gc.ca/Prizes-Prix/Herzberg-Herzberg/Profiles-Profils_fra.asp?ID=1003
	S	http://www.nserc-crsng.gc.ca/Prizes-Prix/Excellence-Excellence/Profiles-Profils_eng.asp?ID=1008	http://www.nserc-crsng.gc.ca/Prizes-Prix/Excellence-Excellence/Profiles-Profils_fra.asp?ID=1008
12	T	http://www.lalettrediplomatique.fr/contribution.php?choixlang=2&id=9&idrub=12	http://www.lalettrediplomatique.fr/contribution.php?id=9&idrub=12
	S	http://www.lalettrediplomatique.fr/contribution.php?choixlang=2&id=9&idrub=12	http://www.lalettrediplomatique.fr/contribution.php?choixlang=1&id=9&idrub=12
13	T	http://www.ledindon.com/en/anti-stress/index.php	http://www.ledindon.com/anti-stress/index.php
	S	http://www.ledindon.com/en/anti-stress/index.php?s=2	http://www.ledindon.com/anti-stress/index.php?s=2
14	T	http://www.lupusae.com/en/a_r2.htm	http://www.lupusae.com/en/a_f_r2.htm
	S	http://www.lupusae.com/cn/c_a_r2.htm	http://www.lupusae.com/en/a_f_r2.htm

Table 1: Examples of missed test pairs (T) and equivalent pairs proposed by our system (S). Numbers in parentheses next to (S) refer to the number of equivalent pairs proposed by our system for a specific webdomain. The URLs are those extracted from the `.lett` files.