Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 152–164

Stand-off Annotation of Web Content as a Legally Safer Alternative to Crawling for Distribution

Mikel L. FORCADA, Miquel ESPLÀ-GOMIS, Juan Antonio PÉREZ-ORTIZ

Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant, E-03071 Alacant, Spain

{mlf,mespla,japerez}@ua.es

Abstract. Sentence-aligned web-crawled parallel text or *bitext* is frequently used to train statistical machine translation systems. To that end, web-crawled sentence-aligned bitext sets are sometimes made publicly available and distributed by translation technologies practitioners. Contrary to what may be commonly believed, distribution of web-crawled text is far from being free from legal implications, and may sometimes actually violate the usage restrictions. As the distribution and availability of sentence-aligned bitext is key to the development of statistical machine translation systems, this paper proposes an alternative: instead of copying and distributing copies of web content in the form of sentence-aligned bitext, one could distribute a legally safer *stand-off annotation* of web content, that is, files that identify where the aligned sentences are, so that end users can use this annotation to privately recrawl the bitexts. The paper describes and discusses the legal and technical aspects of this proposal, and outlines an implementation.

Keywords: bitext, parallel text, stand-off annotation, legal issues, statistical machine translation

1 The importance of sentence-aligned crawled bitext

The importance of *bitext* or *parallel text* in current translation technologies is hard to emphasize. Isabelle et al. (1993) —but also Simard et al. (1993)— are famously quoted for saying that "Existing translations contain more solutions to more translation problems than any other currently available resource", but the formulation of the concept of bitext as a translation object can be traced back to Harris (1988).

For bitexts to be used in two key translation technologies, namely *corpus-based machine translation* —particularly *statistical machine translation* (Koehn, 2009), but also *example-based machine translation* (Carl and Way, 2003)— and *computer-aided translation* (Bowker and Fisher, 2010), they have to be segmented and aligned, usually sentence by sentence. Sentence-aligned bitexts, frequently in the form of *translation memories*, are usually obtained as a by-product of computer-aided translation processes, and many of them have been made publicly available, such as DGT-MT, the translation memory of the European Commission's Directorate General for Translation (Steinberger et al., 2012); a comprehensive repository of such sentence-aligned bitexts is provided by OPUS[1] (Tiedemann, 2012).

But in view of the fact that the Internet is packed with webpages which are mutual translations, it is not uncommon for researchers and practitioners to build sentence-aligned bitext by harvesting these webpages, pairing them, sentence-aligning them, and making the resulting corpora publicly available. The most famous example would probably be the Europarl corpus (Koehn, 2005).

Contrary to what may be commonly believed, distribution of web-crawled bitext is far from being free from legal implications,[2] and may sometimes actually violate the usage restrictions of web content, as will be discussed in Section 2. As the distribution and availability of sentence-aligned bitext is key to the development of statistical machine translation systems —in particular when it comes to adapt an existing system to a specific domain (Pecina et al., 2012)—but also to save professional translation effort, Section 3 proposes an alternative: instead of copying and distributing copies of web content in the form of sentence-aligned bitext, one could distribute a legally safer *stand-off annotation* of web content, that is, files that identify where the aligned sentences are, so that end users can use software and this annotation to privately or locally recrawl the bitexts they need. Section 4 surveys related standards and technologies, and an implementation is sketched in Section 5. Concluding remarks (Section 6) end the paper.

2 Legal problems

Considering that a sentence-aligned bitext is an example of the general concept of *corpus*, and that web-crawling is an example of *compiling*, the statement by Baker et al. (2006), p. 48, is clearly pertinent, even if obvious: "Corpus compilers need to observe copyright law by ensuring that they seek permission from the relevant copyright holders to include particular texts. This can only be a difficult and time-consuming process as copyright ownership is not always clear [...]. If the corpus is likely to be made publicly available, copyright holders may require a fee for allowing their text(s) to be included".

One might think that web content is not subject to copyright, but this is seldom the case. On the one hand, some web content has explicitly stated licenses which may impact on products derived from it. For instance, Wikipedia[3] uses the Creative Commons Attribution-Sharealike license,[4] which is quite open about the reuse of content, but requires all derivatives to carry the same license. Web-based newspapers usually have more restrictive terms: for instance, the web edition of *The New York Times* uses a typical copyright notice: "You may not modify, publish, transmit, participate in the transfer or sale of, reproduce [...], create new works from, distribute, perform, display, or for

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

[2] Many parallel corpora crawled from the Internet are distributed disregarding the copyright on the original documents from which they were extracted. A clear example is the case of the Europarl corpus for which authors claim (see `http://www.statmt.org/europarl/`) that: *we are not aware of any copyright restrictions of the material.*

[3] `https://www.wikipedia.org/`

[4] `https://creativecommons.org/licenses/by-sa/3.0/`

any way exploit, any of the Content [...] in whole or in part."[5] In another example, participants in the Microblog Track of the Text Retrieval Conference (TREC) interact with a corpus of tweets stored remotely through a search API since 2013. The motivation behind this arrangement —as opposed to the one used in former editions, where the corpus could be downloaded— is to adhere to Twitter's terms of service as they "forbid redistribution of tweets, and thus it would not be permissible for an organization to host a collection of tweets for download" (Lin and Efron, 2013).

Note that usage rights management in the case of bitext corpora compiled from various sources with different licenses may be very complex, which would be particularly hard for non-experts. But what happens when web content is provided without an explicit copyright statement? One would think that it might be possible to use it freely, but this is not the case. According to customary interpretations of the Berne Convention,[6] the most important international agreement dealing with copyright joined by 170 states, copyright notices are optional, works are automatically copyrighted when they are created, and, by default, this means that acts of copying, distribution or adaptation without the author's consent are forbidden. Therefore, in most countries, copyright is automatic and "all rights reserved". The Berne Convention, as an international agreement, may not take into account the variations that copyright law may have in each country.[7] However, it authorizes countries to allow a *fair use* of copyrighted works. In line with this, the Copyright Directive of the European Union[8] states that:

> "Member States may provide for exceptions or limitations to the rights [...] in the following cases: [...] use for the sole purpose of illustration for teaching or scientific research [...] and to the extent justified by the non-commercial purpose to be achieved" (Article 5.3).

In the UK, for instance, there is a prominent *exception to copyright* dealing with text and data mining for non-commercial purposes,[9] which does not exist in other countries. Along these lines, the European Commission recently[10] outlined its vision to modernise European Union copyright rules in order to "make it easier for researchers to use text and data mining technologies to analyse large sets of data"; note, however, that corpus redistribution may still face a lot of risks and uncertainties.

All this means that, depending on the copyright terms of the source material, web-crawled bitexts may not be freely distributed. Tsiavos et al. (2014) discuss in detail the legal issues involved in the distribution of web-crawled data, and even give a number of worked examples. Two main conclusions are:

[5] `http://www.nytimes.com/content/help/rights/terms/terms-of-service.html`

[6] Berne Convention for the Protection of Literary and Artistic Works, 9 September 1886, as last revised at Paris on 24 July 1971, 1161 U.N.T.S. 30.

[7] "Copyright law is not fully harmonized at the international level and, hence, it is extremely difficult to provide a generic answer for the entirety of the situations involving more than one jurisdiction, where possible act of infringement takes place." (Arranz et al., 2013)

[8] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001.

[9] `https://www.gov.uk/guidance/exceptions-to-copyright`

[10] `http://europa.eu/rapid/press-release_IP-15-6261_en.htm`

- In general, publish only after clearing copyright of the content with the holder (if all of the crawled content has the same public license and it allows redistribution under specific terms, one can of course avoid clearing copyright).
- Abide by a *notice and take down* procedure,[11] much in the way in which online hosts remove content following notice such as court orders or allegations that content infringes copyright.

Also, they suggest that if one cannot clear copyright, it may be safer to publish a derivative of the crawled content from which it is impossible to reconstruct the original source. In fact, when discussing *annotations* as a special case of derivative works, Tsiavos et al. (2014, p. 41) conclude that "unless [the annotations] reproduce part of the original work they do not constitute a problem". Similarly, Arranz et al. (2013) analyse the legal status of different acts involving web crawling of data and web services built around them, and state that "if what is communicated to the public is the actual data either in their original or their derivative form, then this constitutes yet another act restricted by copyright law. If, however, the end user is only the recipient of a web service that implements the web crawling and processing without any direct communication of the actual web data, then copyright law is not activated at all".

It is in this context that avoiding redistribution and moving usage rights management to the final user shows its advantages: as content is not republished but referred to, there is no need to handle copyright, and use after recrawling *chez* the end-user is more likely to be considered fair use.

3 The proposal: stand-off annotation

Following the rationale in the previous section, it is proposed that instead of publicly distributing web-crawled sentence-aligned bitexts, a *stand-off annotation* of the Internet will be distributed, an annotation detailed enough for the end user to efficiently *recrawl* locally the sentence-aligned bitext using appropriate software, on the grounds that an annotation cannot be considered a derived work but rather a description of existing content geared at a specific purpose, not too different from the concepts of *metadata* or *bibliographical reference* as used in scholarly publishing. Public distribution is avoided, and, as a result, the need to clear copyright disappears altogether for corpus compilers, and the responsibility of rights management is passed on to the end user.

Many of the usages by *end users* could actually fall into what is called *fair use*: for instance, a translator may use and modify selected segments of a web-crawled translation memory to produce the translation of a new document. The legal status of more extensive usages such as when a web-crawled sentence-aligned bitext is used to train or domain-adapt a statistical machine translation is less clear, but some machine translation systems available on the web (Google Translate[12] and Bing Translator[13]) rely in part on web-crawled content[14] and this usage, to the best of our knowledge, has not been the subject of any solid legal challenge.

[11] `https://en.wikipedia.org/wiki/Notice_and_take_down`

[12] `http://translate.google.com`

[13] `https://www.bing.com/translator/`

[14] `http://v.gd/tausgt` (shortened URL)

The Text Encoding Initiative[15] defines “Stand-off markup (also known as remote markup or stand-off annotation)” as “the kind of markup that resides in a location different from the location of the data being described by it. It is thus the opposite of inline markup, where data and annotations are intermingled within a single location”.

The idea of stand-off annotation of corpora is not new, but to the best of our knowledge, it has not been used before to directly annotate web content *at large*, that is, *in the wild*. However, there are some examples of stand-off annotation for building bitexts from collections of documents, as it is the case of the JRC-Acquis (Steinberger et al., 2006) corpus, which is distributed as a collection of monolingual documents and a stand-off annotation file that describes the segment-aligned bitexts that can be obtained for every pair of languages with different alignment tools. In this case, this stand-off annotation is rather simple, given that the monolingual documents are preprocessed so every segment of the text is identified with a code that is later used to relate parallel segments across bitexts. This is, in fact, the usual stand-off approach to corpus annotation, where some auxiliary inline annotation is involved:

> “A middle course is for the original corpus publication to have a scheme for identifying any sub-part. Each sentence, tree, or lexical entry, could have a globally unique identifier, and each token, node or field (respectively) could have a relative offset. Annotations, including segmentations, could reference the source using this identifier scheme (a method which is known as stand-off annotation). This way, new annotations could be distributed independently of the source, and multiple independent annotations of the same source could be compared and updated without touching the source.” (Bird et al., 2009, ch. 11).

We could call this *impure* stand-off, as the object being annotated has to be segmented and provided with identifiers. As this is not possible with read-only web content at large, we have to resort to *pure* stand-off annotation, as described below. The following proposals for *crawled bitext* and *crawled translation memory* are based on the concept of *stand-off annotation* of the web as it is found at the time of crawling.

3.1 Deferred bitext crawl

The core of the proposal for crawled bitext, which will be called a *deferred bitext crawl* is *a pair of uniform resource identifiers (URIs)*, one pointing at the *left document*, and another one pointing at the *right document*, such that they are selected as being mutual translations at the time of crawling. To the pair of URIs, one has to add some metadata:

- The *date and time* of annotation.
- The *languages* of the two texts, each one with an optional indicator of how confident the annotating crawler is that they are actually written in those languages.
- *Checksum information* for both the left and right documents, that will be used to ensure that the texts have not changed since they were crawled. Note that while checksum information could be weakly considered as a derivative, it does not allow the reconstruction the original content: it would have to be recrawled.

[15] `http://wiki.tei-c.org/index.php/Stand-off_markup`

- Optionally, one or more indicators expressing the *confidence* with which the two texts are taken to be mutual translations.

This information may be used to recrawl the two sides of the bitexts and check that they have not changed since they were crawled and classified as being a bitext. Those bitexts not passing the test should be discarded.[16]

3.2 Deferred translation memory crawl

A product that could be derived by selecting sentence pairs from a set of *deferred bitext crawls*, after aligning their sentences, is the *deferred sentence-aligned bitext crawl* (also *deferred translation memory crawl* or *deferred training corpus crawl*): a set (not necessarily ordered) of sentence pairs, each one completely independent, in which every pair is described by:

- The *date and time* of annotation.
- The *languages* of the two sentences, each one with an optional confidence indicator.
- The URI of the file from which each sentence is taken.
- A record indicating the location of each sentence, such as the position of the first character of the sentence, and either the position of the last character or the length in characters of the sentence.
- The checksum value (or other values that ease *integrity check*) at annotation time.
- Optionally, one or more indicators expressing the *confidence* with which the two sentences are taken to be mutual translations (derived from the bitext confidences above, but optionally refined for this specific pair of sentences).

4 Relevant standards and technologies

This paper does not aim at proposing a final solution, but rather at trying to convince the reader that existing technologies may make the sketch in Section 3 technically feasible by actually advancing the main features of the solution. To that end, a survey of related standards and technologies is provided in this section. The main technical requirement is to have *locators* that allow us to point at specific fragments in an HTML document.[17] Ideally, these locators should be sufficiently specific so that changes in the original document can be detected and, in addition to this, error recovery strategies could be implemented in order to find the segment in a different location.

4.1 Integrity checks

The W3C Web Annotation Working Group launched in 2014 with the aim of developing a set of recommendations for web annotation, which will include specifications regarding

[16] "It is better to cause stand-off annotations to break on such components of the new version than to silently allow [them] to refer to incorrect locations." (Bird et al., 2009, ch 11).

[17] All of the discussion in this paper assumes that webpage content will be in HTML, some XML-based text format, and in some cases plain text: an extension to deal with PDF or wordprocessor documents published in websites falls out of the scope of this paper.

robust anchoring into third-party documents. Robustness against modifications in the URL, in the content text or in the underlying structure of the HTML document is an important feature for the systems processing this kind of locators. A common solution is to extend the locator with information about the matched text along with some of the text immediately before and after it,[18] but this practice could lead to copyright infringement. A more covenient option would be in that case to rely on character positions.[19]

There is also a plethora of message-digest and checksum algorithms that may be used to detect changes in the segments pointed at by the stand-off annotation in the deferred crawls described in Section 3. In addition to the MD5[20] message digests,[21] there are alternatives such as SHA-2:[22] most have publicly available implementations.

Link death is obviously a major issue here. A number of studies have analysed the persistence of URLs over time: Gomes and Silva (2006) found that the lifetime of URLs follows a logarithmic distribution in which only a minority persists for periods longer than a few months; Lawrence et al. (2001) studied a database of computer science papers and found that around 30-40% of links were broken, but they could manually found the new location of the page (or highly related information) 80% of the times. In fact, solutions to find the new location of the content when it has been moved, have been proposed ranging from the use of *uniform resource names* (URNs)[23] to heuristic strategies for automatic fixing of dead links (Morishima et al., 2009). Park et al. (2004) found that a lexical signature consisting of several key words is usually sufficient to obtain the new location, which suggests that these key words could be incorporated into the extended locators proposed in our paper. A different, more limited[24] approach (Resnik and Smith, 2003) crawls only non-volatile resources such as the Internet Archive.[25]

4.2 Linking to a fragment of a document

In the definition of URI,[26] the only provision to refer to parts of a webpage occurs through the use of fragment identifiers using the symbol "#", as in the example: `http://server.info/folder/page.html#section2`; however, this presumes the existence of identified anchors in the HTML document. A standard that could be repurposed to

[18] See `https://w3c.github.io/web-annotation/model/wd/#text-quote-selector` or `https://hypothes.is/blog/fuzzy-anchoring/`.

[19] The project Emphasis by The New York Times (see `http://open.blogs.nytimes.com/2011/01/11/emphasis-update-and-source/`) uses keys made up of the first characters from the first and last words in the segment, which constitutes a more compact description and avoids the need to copy text verbatim.

[20] `https://en.wikipedia.org/wiki/MD5`

[21] `https://www.ietf.org/rfc/rfc1321.txt`

[22] `https://en.wikipedia.org/wiki/SHA-2`

[23] `https://www.w3.org/TR/uri-clarification/`

[24] Even though it is possible to use this repository for a more stable version of some contents, it is worth noting that: (a) it does not cover every website on the Internet, and (b) the websites stored in the Internet Archive are not continuously crawled, which means that some live contents may not be available until a new crawl is carried out.

[25] `https://archive.org/`

[26] RFC 3986, `https://www.ietf.org/rfc/rfc3986.txt`.

refer to specific character offsets in a webpage is RFC 5147,[27] "text/plain fragment identifiers", which however deals only with content of the *text/plain* media type, but not with *text/html* which would be the usual media type for webpage content.[28] Note that RFC 5147 already provides the means to implement *integrity checks* and explicitly supports the MD5 message digest standard.[29]

While RFC 5147 could be repurposed for general web content, it does not take into account the structure of the document; indeed, most edits to a webpage usually occur in a way that its structure is only modified locally. Using character offsets would mean that all text after each single edit could fail the integrity check and therefore be discarded: a structure-aware approach could be beneficial to avoid such massive losses of content, the closest candidates being:

- *XPointer*,[30] a system to address components of an XML document, can only be applied to valid XML documents and most webpages are not (they would have to be univocally transformed or normalized into valid XML documents, and pointing would be through the intermediate normalized document). Specific characters inside the contents of an XML element can be linked via the `substring` function.
- Cascaded style sheet (CSS) selectors,[31] used to provide a presentation for an HTML document,[32] do not require it to be a valid XML document; they can therefore operate on a wider range of webpages but they cannot address specific characters. There is some interest in extending the standard in this direction,[33] and indeed extensions to address specific letters[34] have been implemented as JavaScript libraries.
- Canonical Fragment Identifier for EPUB,[35] a method for referencing arbitrary content within electronic books in EPUB format (a format based on HTML). Its linking notation uses a combination of *child sequences*[36] (similar to those defined in XPointer with the `element` scheme) and anchor identifiers, but it is not as robust and expressive as CSS selectors or XPointer. It also allows for character offsets in the form of ranges such as `2:5`.

A combination of one or more of the mentioned standards could form the basis for specifying locators that could be used to point at any character span in the web.

[27] `https://tools.ietf.org/rfc/rfc5147.txt`

[28] RFC 7111 (`https://tools.ietf.org/rfc/rfc7111.txt`) provides fragment identifiers for the *text/csv* media type.

[29] See also the work by Hellmann et al. (2012) for more character-level proposals.

[30] `https://www.w3.org/TR/xptr-xpointer/`

[31] `https://www.w3.org/TR/css3-selectors/`

[32] CSS selectors are also used to point at elements in the document in JavaScript.

[33] `https://css-tricks.com/a-call-for-nth-everything/`

[34] `http://letteringjs.com/`

[35] `http://www.idpf.org/epub/linking/cfi/epub-cfi.html`

[36] An example of a child sequence is `3/1` which represents the second child (counts start at zero) of an element that is the fourth child in the current context.

4.3 Leveraging TMX

A modified version of TMX, the *translation memory exchange* format[37] could be used to distribute *deferred training corpus crawls* —also called *deferred translation memory crawls*— (see section 3.2); this would allow an easy conversion into TMX —basically by retrieving the content pointed at—, ready for use as a translation memory in most computer-aided translation software; converting them to training corpora for statistical machine translation would also be quite simple and could leverage existing software to do so. The main change would affect the `seg` (segment) element, which would have to be substituted by a stand-off annotation of the segment, which could be called `webseg`, and which would contain the URL of the source document and a specification of the actual fragment inside the document; integrity check information could be either added directly to this `webseg` element or as a property using the standard `prop` element. As regards date and time, TMX already supports this information as a property of each translation unit. To avoid repeating URLs in `websegs`, the `header` could contain an element assigning an identifier to each unique source document.

4.4 An example of the TMX-inspired format

Figure 1 illustrates how the TMX format could be transformed into an XML format capable of representing *deferred sentence-aligned bitext crawls* and *deferred translation memory crawls*. This file contains a single sentence pair or *translation unit* (`tu`), having two *variants* (`tuv`), one in English and another one in Spanish (the actual texts are *About the UA* and *Sobre la UA*). A *properties* element (`prop`) in each *variant* contains the MD5 checksum of the text. Instead of using the standard TMX *segment* element (`seg`), a *web segment* element (`webseg`) contains a pointer to a particular segment, made up of an URL, a fragment identifier using Xpointer notation, and a character range inside the selected element (`0:11` in English and `0:10` in Spanish).

5 Implementation: stand-off crawlers

Given the fact that there is a number of bilingual web crawlers able to harvest bitexts from the Internet, such as Bitextor (Esplà-Gomis and Forcada, 2010), ILSP Focused Crawler (Mastropavlos and Papavassiliou, 2011), STRAND (Resnik and Smith, 2003), BITS (Ma and Liberman, 1999), or WeBiText (Désilets et al., 2008), it seems more reasonable to consider adapting an existing parallel data crawler to produce deferred translation memories than implementing a new stand-off crawler from scratch. In general, most of these parallel data crawlers work following a similar process:

1. several documents from a given website are downloaded;
2. documents are pre-processed and their language is identified;
3. parallel documents are identified (document alignment) using heuristics;
4. optionally, parallel documents are segment-aligned.

[37] `https://www.gala-global.org/tmx-14b`

```
<?xml version="1.0" encoding="UTF-8"?>
<tmx version="1.4">
  <header creationtool="Deferred Corpus Creator"
      creationtoolversion="0.95"
      datatype="text/html" segtype="sentence"
      adminlang="en" srclang="en" o-tmf="web"/>
  <body>
    <tu tuid="1">
      <prop type="x-alignment_confidence">0.86</prop>
      <tuv xml:lang="en" date="20161105T153005Z">
        <prop type="x-lang_confidence">0.91</prop>
        <prop type="x-md5">
          28709ee845d8efaf62318210ecd8ca82
        </prop>
        <webseg>
         http://web.ua.es/en/about-the-ua.html#fragment(//*[@id=&
             quot;parteSuperiorPagina&quot;]/div/h1/0:11)
        </webseg>
      </tuv>
      <tuv xml:lang="es" date="20161105T153013Z">
        <prop type="x-lang_confidence">0.73</prop>
        <prop type="x-md5">
          d502972dbfc178f2c1085875890c2144
        </prop>
        <webseg>
         http://web.ua.es/va/sobre-la-ua.html#fragment(//*[@id=&
             quot;parteSuperiorPagina&quot;]/div/h1/0:10)
        </webseg>
      </tuv>
    </tu>
  </body>
</tmx>
```

Fig. 1. Example of a deferred translation memory crawl containing a single translation unit (see text for details).

Therefore, the problems faced when adapting any parallel data crawler to the purposes of our work would be similar in any of them. This section discusses how these crawlers could be adapted to produce deferred translation memories (Section 4.3).

One of the main obstacles to adapt a state-of-the-art parallel data crawler for the purpose of our work is that, in most of the cases, they do not obtain the translation memories directly from the original documents downloaded from the web: these documents are pre-processed before segment-aligning them. For example, Bitextor and ILSP Focused Crawler normalise HTML documents into XHTML by using the tool *Apache Tika*,[38] and remove boilerplates with the tool *Boilerpipe*.[39] In addition, most crawlers remove the HTML mark-up before segment alignment. This means that both the HTML structure and the content of the documents may be modified before obtaining the final segment alignment. To deal with this problem it would be necessary to annotate the text in the document with the reference of its position in the original document. This could be done by using additional HTML mark-up, which would be preserved during pre-processing.

After document alignment and HTML mark-up cleaning, every document would consist of a collection of text blocks for which their current offset is mapped to their

[38] `http://tika.apache.org/`

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

position in the original document. At this point, sentence splitting is carried out, which yields several segments from a single text block for which its position in the original document is known. It will therefore be necessary to obtain the position of every segment in the original document, which should be straightforward knowing that every text block appears in a known position of the HTML tree in the original document. In this case, it is sufficient to keep track of the offset of the first and last characters of the segments obtained taking the position identifier of the original document as a reference.

By adapting existing parallel data crawlers to keep track of the processing carried out to transform the original documents to the final segment-aligned parallel corpus, a TMX-like document such as the one described in Section 4.3 could be obtained by replacing the actual sentence pairs obtained after sentence alignment by the mapping to their original locations.

6 Concluding remarks

This paper has laid the foundations and advanced a proposal for a new way to distribute web-crawled sentence-aligned bitext to avoid legal problems associated to distribution. The main idea is to distribute a stand-off annotation of the *wild* web content that makes up the aligned sentences or translation units, which is called a *deferred translation memory crawl* or *deferred training corpus crawl*. It is proposed that a modification of the existing TMX standard for translation memories is used as the basis of the new standoff format. This makes it easy to modify existing crawlers such as Bitextor and ILSP Focused Crawler to produce this kind of output. Although in this paper a tentative syntax to point at the linked segments has been outlined, it could change and evolve as specifications regarding robust anchoring to third-party documents are developed by the recently created W3C Web Annotation Working Group. If the proposal in this paper is adopted, we could be looking at massive repositories of deferred translation memories that could be legally distributed without having to manage the copyright of the original content, and which could be used by end users (professional translators, statistical machine translation practitioners) to recrawl the web and use the selected content under *fair use* provisions.

Acknowledgements

Funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran) is acknowledged. The authors would like to thank the anonymous reviewers for their valuable suggestions.

References

Victoria Arranz, Khalid Choukri, Olivier Hamon, Núria Bel, and Prodromos Tsiavos. PANACEA project deliverable 2.4, annex 1: Issues related to data crawling and licensing. http://cordis.europa.eu/docs/projects/cnect/4/248064/080/deliverables/001-PANACEAD24annex1.pdf, 2013.

Paul Baker, Andrew Hardie, and Tony McEnery. *A glossary of corpus linguistics*. Edinburgh University Press, 2006.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009. `http://www.nltk.org/book/ch11.html`.

Lynne Bowker and Des Fisher. Computer-aided translation. *Handbook of Translation Studies*, 1: 60, 2010.

Michael Carl and Andy Way. *Recent advances in example-based machine translation*, volume 21. Springer Science & Business Media, 2003.

Alain Désilets, Benoit Farley, M Stojanovic, and G Patenaude. WeBiText: Building large heterogeneous translation memories from parallel web content. In *Proceedings of Translating and the Computer*, pages 27–28, London, UK, 2008.

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

Daniel Gomes and Mário J. Silva. Modelling information persistence on the web. In *Proceedings of the 6th International Conference on Web Engineering*, ICWE '06, 2006.

Brian Harris. Bi-text, a new concept in translation theory. *Language Monthly*, 54:8–10, 1988.

Sebastian Hellmann, Jens Lehmann, and Sören Auer. Linked-data aware URI schemes for referencing text fragments. In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management*, pages 175–184, Galway City, Ireland, 2012.

Pierre Isabelle, Marc Dymetman, George Foster, Jean-Marc Jutras, Elliott Macklovitch, François Perrault, Xiaobo Ren, and Michel Simard. Translation analysis and translation automation. In *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*, pages 1133–1147. IBM Press, 1993.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005.

Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

Steve Lawrence, David M. Pennock, Gary William Flake, Robert Krovetz, Frans M. Coetzee, Eric Glover, Finn Årup Nielsen, Andries Kruger, and C. Lee Giles. Persistence of web references in scientific research. *Computer*, 34(2):26–31, 2001.

J. Lin and M. Efron. Overview of the TREC-2013 Microblog Track. In *Proceedings of the Twenty-Second Text REtrieval Conference (TREC 2013)*, 2013.

Xiaoyi Ma and Mark Liberman. Bits: A method for bilingual text search over the web. In *Machine Translation Summit VII*, pages 538–542, Singapore, Singapore, 1999.

Nikos Mastropavlos and Vassilis Papavassiliou. Automatic acquisition of bilingual language resources. In *Proceedings of the 10th International Conference of Greek Linguistics*, 2011.

Atsuyuki Morishima, Akiyoshi Nakamizo, Toshinari Iida, Shigeo Sugimoto, and Hiroyuki Kitagawa. Bringing your dead links back to life: A comprehensive approach and lessons learned. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, 2009.

Seung-Taek Park, David M. Pennock, C. Lee Giles, and Robert Krovetz. Analysis of lexical signatures for improving information persistence on the world wide web. *ACM Trans. Inf. Syst.*, 22(4):540–572, 2004.

Pavel Pecina, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, Josef Van Genabith, and RIC Athena. Domain adaptation of statistical machine translation using web-crawled resources: a case study. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 145–152, 2012.

Philip Resnik and Noah A. Smith. The Web as a parallel corpus. *Computational Linguistics*, 29 (3):349–380, 2003.

Michel Simard, George F. Foster, and François Perrault. Transsearch: A bilingual concordance tool. *Centre d'innovation en technologies de l'information, Laval, Canada*, 1993.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, and D. Tufiş. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 2142–2147, Genoa, Italy, 2006.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, 2012.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *LREC*, pages 2214–2218, 2012.

Prodromos Tsiavos, Stelios Piperidis, Maria Gavrilidou, Penny Labropoulou, and Tasos Patrikakos. Qtlaunchpad public deliverable d4.5.1: Legal framework. `http://www.qt21.eu/launchpad/system/files/deliverables/QTLP-Deliverable-4_5_1_0.pdf`, 2014.

Received May 2, 2016 , accepted May 9, 2016