


Stand-off annotation of web content as a legally safer alternative to bitext crawling for distribution¹

Mikel L. Forcada **Miquel Esplà-Gomis** Juan A. Pérez-Ortiz
`{mlf,mespla,japerez}@dlsi.ua.es`

Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03690 Sant Vicent del Raspeig, Spain

May 31, 2016

¹Work funded by the EC through PIAP-GA-2012-324414. **Abu-Matran** 

Outline

- 1 Motivation
- 2 Deferred corpora
- 3 Our proposal for deferred translation memories
- 4 Implementation in existing parallel data crawlers

Outline

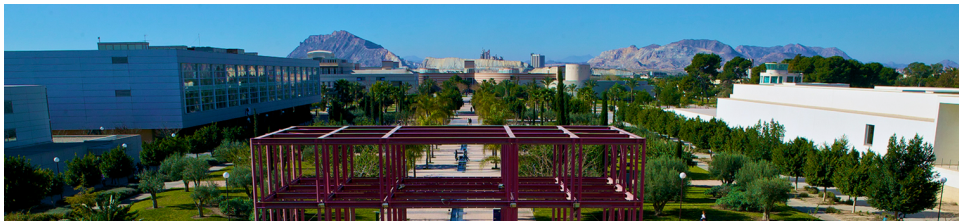
- 1 Motivation
- 2 Deferred corpora
- 3 Our proposal for deferred translation memories
- 4 Implementation in existing parallel data crawlers

Parallel corpora are a useful resource!

- Relevant resource for many NLP problems, especially for **statistical machine translation (SMT)**
- Many parallel corpora are available **on the Internet**:
 - ▶ **In the wild**, waiting to be harvested:
 - ★ news, official documents, technical documentation, customer support, public service information, etc.
 - ▶ **Already packaged**, usually sentence-aligned:
 - ★ Europarl, MultiUN, EMEA, News Commentary, SETIMES, TED talks, etc.
 - ★ Many of these available in OPUS.

Parallel contents on the Web

<http://web.ua.es/va/sobre-la-ua.html>



La Universitat

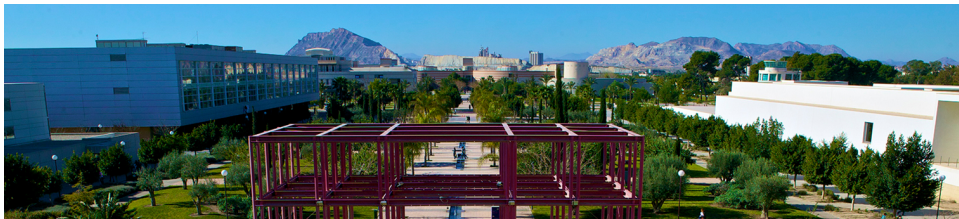
SOBRE LA UA

Missatge de benvinguda

Servisquen aquestes pàgines web que esteu visitant per a donar-vos la nostra càlida benvinguda virtual a la Universitat d'Alacant i agrair el vostre interès a conèixer-nos millor. Hi trobareu descrita l'estructura i la rica diversitat de les activitats que desenvolupa la nostra Universitat, com també les línies amb les quals aspirem a traçar el futur d'aquesta institució, esforçant-nos a fer-la cada dia més profitosa per a les necessitats i els anhels de la societat.

Parallel contents on the Web

<http://web.ua.es/en/about-the-ua.html>



The University

ABOUT THE UA

Welcome message

We are pleased to give you our warmest welcome to the Institutional Website of the University of Alicante and we would like to thank for your interest in getting to know us better. Here you can find the description of its structure and the wide range of activities that can be carried out at our University, as well as the lines with which we plan to trace the future of our Institution, making an effort in doing our best to cover the needs and desires of society.

Example: standard TMX

```
<tmx version="1.4">
  <header .../>
  <body>
    <tu tuid="1">
      <tuv xml:lang="en" date="20161105T153005Z">
        <seg>About the UA</seg>
      </tuv>
      <tuv xml:lang="ca" date="20161105T153013Z">
        <seg>Sobre la UA</seg>
      </tuv>
    </tu>
    ...
  </body>
</tmx>
```



- Online content is usually protected by **restrictive copyright**
 - ▶ *Standard interpretation* of the **Berne Convention**: automatic copyright, all rights reserved.
- Some current solutions to distribute corpora based on public text are:
 - ▶ Declaring that copyright terms are **unknown**
 - ▶ Removing documents **under request** (“notice and take down”)
 - ▶ **Process** data before distribution to difficult to guess its source (randomize, obtain sub-products, etc.)
- These solutions may violate the **Berne Convention**

What to do?

- Build parallel corpora only on **authorised resources** (Wikipedia, free/open-source software, etc.)?
- **Request permission from** of every resource we want to use?
- **Forget about distributing** parallel corpora crawled from the Internet?
- Keep distributing data (and **hope you will not be sued**)?



Outline

- 1 Motivation
- 2 Deferred corpora
- 3 Our proposal for deferred translation memories
- 4 Implementation in existing parallel data crawlers

What do we propose?

- Using a **stand-off annotation** of parallel data *in the wild* on the web
...
- ... that allows to **easily re-download** (*re-crawl*) a parallel corpus ...
- ... transferring usage restrictions to the **final user**.

Standoff annotation #1: The *deferred bitext crawl*

Deferred bitext crawl: Parallel corpus consisting of pairs of identifiers of documents which are parallel; data needed:

- URLs of both documents
- languages of both documents
- checksum codes (to check integrity) for every pair of documents
- an indication about the confidence that the documents are mutual translations

Standoff annotation #2: The *deferred translation memory crawl*

Deferred translation memory crawl: Parallel corpus consisting of identifiers of segment pairs which are parallel; data needed:

- URLs of the documents the segments come from
- pair of languages
- positions and lengths of the segments in each document
- checksum codes (to check integrity) for every pair of segments
- an indication of the confidence that the pair of segments are mutual translation

Pros and cons

	Deferred corpus crawl	Deferred translation memory crawl
pros	<p>simpler data</p> <p>smaller corpus (less data)</p>	<p>more robust to changes (integrity checks discard only some segment pairs)</p> <p>easier to re-build the corpus (no need to re-align documents)</p>
cons	<p>need to re-build the corpus</p> <p>less robust to changes (if checksum changes, the whole document pair is discarded)</p>	<p>need to re-build the corpus</p> <p>larger and more complex description of the corpus</p>

Requirements

- Annotation convention allowing to identify relations between specific segments of text in documents on the Web
- Semantics rich enough to contain all the information needed to re-build the corpus
- Method robust to changes on the contents of the documents

Existing technologies & standards

- Integrity checks: hash functions that provide signatures from data
 - ▶ SHA-2: 224- to 512-byte signature
 - ▶ MD5: 32 hexadecimal characters signature
- Linking to a fragment of a document:
 - ▶ RFC 5147: allows to refer character offsets in a webpage, but only on plain text
 - ▶ XPointer: a system to address components of an XML document
 - ▶ CSS selectors: a system to address components of an HTML (not necessarily XML-valid)
 - ▶ Canonical Fragment Identifier for EPUB: allows referencing arbitrary content within EPUB books (based on HTML); not as robust and expressive as CSS selectors or XPointer
- Translation memories:
 - ▶ TMX: XML-based standard for translation memories

Outline

- 1 Motivation
- 2 Deferred corpora
- 3 Our proposal for deferred translation memories**
- 4 Implementation in existing parallel data crawlers

Example: standard TMX

```
<tmx version="1.4">
  <header .../>
  <body>
    <tu tuid="1">
      <tuv xml:lang="en" date="20161105T153005Z">
        <seg>About the UA</seg>
      </tuv>
      <tuv xml:lang="ca" date="20161105T153013Z">
        <seg>Sobre la UA</seg>
      </tuv>
    </tu>
    ...
  </body>
</tmx>
```

Extending TMX with deferred aligned segments (WebSeg)

```
<tmx version="1.4">
<header .../>
<body>
  <tu tuid="1">
    <tuv xml:lang="en" date="20161105T153005Z">
      <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">
      <webseg>
        http://web.ua.es/va/sobre-la-ua.html#fragment (
          /*[@id=&quot;parteSuperiorPagina&quot;]/div/h1/0:10)
      </webseg>
    </tuv>
  </tu>
</body>
</tmx>
```

Extending TMX with deferred aligned segments (WebSeg)

```
<webseg>  
  http://web.ua.es/en/about-the-ua.html  
  #fragment (  
    /*["@id="parteSuperiorPagina";]  
    /div/h1/0:11  
  )  
</webseg>
```

Final deferred translation memory crawl

```
<tmx version="1.4">
  <header .../>
  <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...</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...</webseg>
      </tuv>
    </tu>
  </body>
</tmx>
```

Final deferred translation memory crawl

```
<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/va/sobre...</webseg>
</tuv>
```

Outline

- 1 Motivation
- 2 Deferred corpora
- 3 Our proposal for deferred translation memories
- 4 Implementation in existing parallel data crawlers

How do parallel data crawlers work? /1

- **Parallel data crawlers** are tools able to build parallel corpora from multilingual websites:
 - ▶ BITS (Ma and Liberman, 1999), or
 - ▶ STRAND (Resnik and Smith, 2003),
 - ▶ Bitextor* (**Esplà-Gomis** and **Forcada**, 2010),
 - ▶ ILSP Focused Crawler* (Mastropavlos and Papavassiliou, 2011),

*free/open-source

How do parallel data crawlers work? /1

- These tools generate translation memories from multilingual contents on the Web by:
 - 1 Downloading documents from a website
 - 2 Pre-processing documents and identifying their language
 - 3 Identifying parallel documents
 - 4 Aligning parallel documents at the segment level
- Is it possible to **adapt them to produce deferred translation memories?**

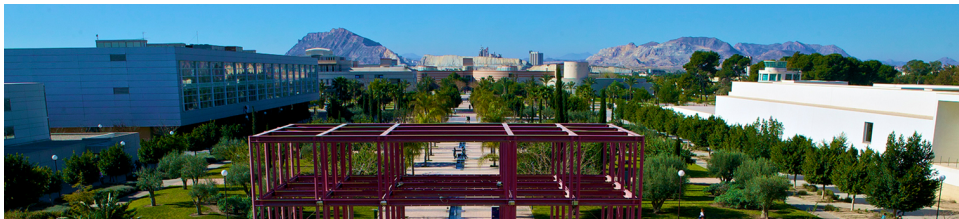
Downloading documents from a website

Actions to adapt a parallel data crawler:

- 1 Downloading documents from a website
 - ▶ Store URL
- 2 Pre-processing documents and identifying their language
- 3 Identifying parallel documents
- 4 Aligning parallel documents at the segment level

Parallel contents on the Web

<http://web.ua.es/en/about-the-ua.html>



The University

ABOUT THE UA

Welcome message

We are pleased to give you our warmest welcome to the Institutional Website of the University of Alicante and we would like to thank for your interest in getting to know us better. Here you can find the description of its structure and the wide range of activities that can be carried out at our University, as well as the lines with which we plan to trace the future of our Institution, making an effort in doing our best to cover the needs and desires of society.

Downloading documents from a website

```
<div class="container container-fluid">
<div class="row" id="parteSuperiorPagina">
<div class="col-md-12"><span class="textoSubtituloPagina "
    data-mce-mark="1">The University</span>
<h1 class="textoTituloPagina">About the UA</h1>
<span data-mce-mark="1">Welcome message</span>
<p class="descripcionPagina tresColumnas">We are pleased to
    give you our warmest welcome to the Institutional Website
    of the University of Alicante and we would like to thank
    for your interest in getting to know us better. Here you
    can find the description of its structure and the wide
    range of activities that can be carried out at our
    University, as well as the lines with which we plan to
    trace the future of our Institution, making an effort in
    doing our best to cover the needs and desires of society.
    ...
    <br /> <br /> Thank you very much.</p>
</div>
</div>
</div>
```

URL=<http://web.ua.es/en/about-the-ua.html>

Pre-processing documents and identifying their language

Actions to adapt a parallel data crawler:

- ① Downloading documents from a website
- ② Pre-processing documents and identifying their language
 - ▶ Store XPointer & Canonical Fragment Identifier for EPUB reference
 - ▶ Keep track of the effect of normalisation (removing boilerplates, useless data, etc.) on document structure.
 - ▶ Store language detected
- ③ Identifying parallel documents
- ④ Aligning parallel documents at the segment level

Pre-processing documents and identifying their language

`The University` `[@id="parteSuperiorPagina"]/div/span/0:14`

`<h1>About the UA</h1>` `[@id="parteSuperiorPagina"]/div/h1/0:12`

`Welcome message` `[@id="mensajeBienvenida"]/0:15`

`<p>We are pleased to give you our warmest welcome to the
Institutional Website of the University of Alicante and we
would like to thank for your interest in getting to know u
better. Here you can find the description of its structur
and the wide range of activities that can be carried out
at our University, as well as the lines with which we plan
to trace the future of our Institution, making an effort i
doing our best to cover the needs and desires of society.
...

 Thank you very much.</p>`

`[@id="parteSuperiorPagina"]/div/p/0:1525`

`xml:lang="en"`

Identifying parallel documents

Actions to adapt a parallel data crawler:

- 1 Downloading documents from a website
- 2 Pre-processing documents and identifying their language
- 3 Identifying parallel documents
 - Store list of pairs of documents
- 4 Aligning parallel documents at the segment level

Aligning parallel documents at the segment level

Actions to adapt a parallel data crawler:

- ➊ Downloading documents from a website
- ➋ Pre-processing documents and identifying their language
- ➌ Identifying parallel documents
- ➍ Aligning parallel documents at the segment level
 - ▶ Clean up XML/HTML mark-up
 - ▶ Update positions to reflect text segmentation
 - ▶ Update Canonical Fragment Identifier for EPUB

Pre-processing documents and identifying their language

The University [@id="parteSuperiorPagina"]/div/span/0:14

About the UA [@id="parteSuperiorPagina"]/div/h1/0:12

Welcome message [@id="mensajeBienvenida"]/0:15

We are pleased to give you [@id="parteSuperiorPagina"]/div/p/0:150
our warmest welcome to the
Institutional Website of the University of Alicante and we
would like to thank for your interest in getting to know us
better.

Here you can find the des- [@id="parteSuperiorPagina"]/div/p/151:465
cription of its structure
and the wide range of activities that can be carried out at
our University, as well as the lines with which we plan
to trace the future of our Institution, making an effort in
doing our best to cover the needs and desires of society.

...

Thank you very much. [@id="parteSuperiorPagina"]/div/p/1505:1525

Concluding remarks

- Many corpora are distributed violating Berne convention
- We propose a format based on TMX to align text fragments on the Internet
- The new annotation basically amounts to saying:
 - ▶ Here's where a text segment is in the wild Internet
 - ▶ Here's where another text segment is
 - ▶ They are mutual translations
 - ▶ Crawl them and use them!
- This annotation format may be modified to adapt it to the standards proposed by the W3C Web Annotation Working Group in the future

These slides are free/open-source

This work may be distributed under the terms of

- the Creative Commons Attribution–Share Alike license:
<http://creativecommons.org/licenses/by-sa/4.0/>
- the GNU GPL v. 3.0 License:
<http://www.gnu.org/licenses/gpl.html>

Dual license! E-mail us to get the sources: mlf@ua.es,
mespla@dlsi.ua.es