Building a Corpus of Multi-lingual and Multi-format International Investment Agreements

Sugisaki, Kyoko; Volk, Martin; Polanco, Rodrigo; Alschner, Wolfgang; Skougarevskiy, Dmitriy

Abstract: In this paper, we present an on-going research project whose aim is to develop a new database of international investment agreements that complements existing endeavors. In particular, this paper describes our efforts to build a standard- ized corpus of multi-lingual and multi-format agreement texts in order to enable re- searchers in the fields of international law and economics systematically investigate investment treaties.

Building a Corpus of Multi-lingual and Multi-format International Investment Agreements

Kyoko SUGISAKI [a,1] , Martin VOLK [a] , Rodrigo POLANCO [b] ,
Wolfgang ALSCHNER [b,c] and Dmitriy SKOUGAREVSKIY [c,d]

[a] *Institute of Computational Linguistics, University of Zurich, Switzerland*
[b] *World Trade Institute, University of Bern, Switzerland*
[c] *Graduate Institute of International and Development Studies, Geneva, Switzerland*
[d] *Institute for the Rule of Law, European University, St. Petersburg, Russia*

Abstract. In this paper, we present an on-going research project whose aim is to develop a new database of international investment agreements that complements existing endeavors. In particular, this paper describes our efforts to build a standardized corpus of multi-lingual and multi-format agreement texts in order to enable researchers in the fields of international law and economics systematically investigate investment treaties.

Keywords. Natural language processing, Annotation of International investment agreements, OCR, Language identification

1. Introduction

International investment agreements (henceforth: IIAs) are treaties between two or more countries that are mainly designed for the "protection and liberalization of foreign investment."[1] Since 1990, the number of IIAs has been growing [1] and worldwide, more than 3300 treaties are in force. These treaties are "a key instrument in the strategies of most countries, in particular developing countries, to attract foreign investment."[2] However, they come in a broad variety of formats and languages. The majority of treaties are written in English. Yet, many IIAs are still exclusively in local languages of the contracting countries. Existing databases contain either only treaties in English,[2] or all in original language, and are incomplete.[3] Retrieving such treaties is often difficult for practitioners and scholars.

[1] All authors have been supported by Swiss Network for International Studies (SNIS) grant 37-740.

[2] The examples are Kluwer Arbitration "BITs" [`http://www.kluwerarbitration.com/CommonUI/BITs-countries.aspx`] and Oxford's "Investment Claims" [`http://oxia.ouplaw.com/`]

[3] The database of UNCTAD [`http://investmentpolicyhub.unctad.org/IIA`] is the most advanced and includes 3523 treaties. However, the database misses 772 treaties (22% of the total)

In our project,[4] we collect IIAs, and make them centrally available in one single language (English) and easily accessible by creating a machine-readable structured database. We have so far processed 1026 international investment agreements in HTML, Microsoft Word, and PDF format in more than 30 languages. A few thousand more are in the pipeline.[5] To bring these heterogeneous agreement texts into one format and one language, we have processed PDF documents through optical character recognition (OCR), translated multi-lingual texts into English, and converted unstructured text data into structured data.[6]

This paper describes the project phases from digitization to XML standardization. We focus on the language technology challenges in using OCR and language identification (Section 2), and introduce domain-specific XML mark-up for international investment treaties and automatic text structure recognition for the XML conversion (Section 3).

2. OCR and Automatic Language Identification

Out of the available source documents, 908 were PDFs. About half of these were not digital born but image scans. Therefore, we processed our PDF documents through a state-of-the-art OCR software.[7]

The conversion of treaties from more than 20 languages posed challenges because the software does not automatically identify the language. We thus first processed all documents without specific language settings. This led to unsatisfactory outputs, because the OCR software uses language-specific dictionaries to guess words.[8] Entering the languages of such a large number of documents manually was not a viable option. We therefore used a *n*-gram-based language identification system [3].[9]

Our PDF data collection, in particular, contains multi-lingual documents that include more than one language at the level of sentences, columns, and pages in a document (cf. Figure 1). To handle these three types of code-switching,[10] we segmented the texts into 15 words and changed the letters to lower case. Then we used these units as input to the language identification system. In this way, multi-lingual texts were processed as a unit, short enough to handle sentence-level code-switching but also long enough to handle n-gram in language identification. To obtain a high level of accuracy of language identification, we used the confidence score (i.e. probability estimate for the predicted language)

[4]The Swiss Network for International Studies project *Diffusion of International Law: A Textual Analysis of International Investment Agreements*: `http://www.snis.ch/project_diffusion-international-law-textual-analysis-international-investment-agreements`

[5]Work on collecting IIAs is still in progress in the project, therefore we can not provide a total number of treaties. Yet we have so far collected 3329 treaties.

[6]By unstructured data, we refer to the fact that the texts are not marked with structure in the source data. The treaties themselves are very structured but without markups not accessible for machines.

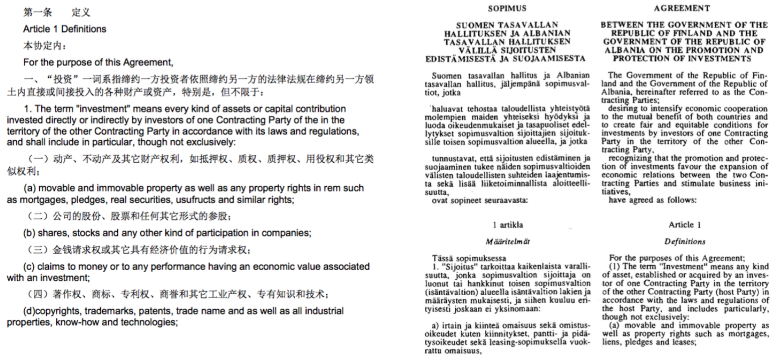
[7]We used Abbyy Recognition Server: `https://www.abbyy.com/recognition-server/`

[8]The problem was also caused, because the OCR system is not restricted to the character set of a specific language.

[9]We used the off-the-shelf langid.py system [`https://pypi.python.org/pypi/langid/1.1.5`]. This tool is suitable to our source legal texts, as the *n*-gram is based on government-related documents among other documents (e.g. Wikipedia and Reuters).

[10]While code-switching usually refers to switching language within a sentence, we use the term in a broader sense.

returned by the language identification system and set a threshold of 0.75. We observed that the language of single-language documents was identified with higher confidence if more than 70% of the units in a document were recognized as the same language. In multi-lingual documents, on the other hand, we found that 25% of the units needed to be identified as one language to provide a good result.[11] Otherwise, we manually examined and corrected the languages of the documents.



第一条　定义

Article 1 Definitions

本协定内：

For the purpose of this Agreement,

一、“投资”一词系指缔约一方投资者依照缔约另一方的法律法规在缔约另一方领土内直接或间接投入的各种财产或资产，特别是，但不限于：

1. The term "investment" means every kind of assets or capital contribution invested directly or indirectly by investors of one Contracting Party of the in the territory of the other Contracting Party in accordance with its laws and regulations, and shall include in particular, though not exclusively:

（一）动产、不动产及其它财产权利，如抵押权、质权、质押权、用役权和其它类似权利；

(a) movable and immovable property as well as any property rights in rem such as mortgages, pledges, real securities, usufructs and similar rights;

（二）公司的股份、股票和任何其它形式的参股；

(b) shares, stocks and any other kind of participation in companies;

（三）金钱请求权或其它具有经济价值的行为请求权；

(c) claims to money or to any performance having an economic value associated with an investment;

（四）著作权、商标、专利权、商誉和其它工业产权、专有知识和技术；

(d)copyrights, trademarks, patents, trade name and as well as all industrial properties, know-how and technologies;

SOPIMUS

SUOMEN TASAVALLAN HALLITUKSEN JA ALBANIAN TASAVALLAN HALLITUKSEN VÄLILLÄ SIJOITUSTEN EDISTÄMISESTÄ JA SUOJAAMISESTA

Suomen tasavallan hallitus ja Albanian tasavallan hallitus, jäljempänä sopimusvaltiot, jotka

haluavat tehostaa taloudellista yhteistyötä molempien maiden yhteiseksi hyödyksi ja luoda oikeudenmukaiset ja tasapuoliset edellytykset sopimusvaltion sijoittajien sijoituksille toisen sopimusvaltion alueella, ja jotka

tunnustavat, että sijoitusten edistäminen ja suojaaminen tukee näiden sopimusvaltioiden välisten taloudellisten suhteiden laajentamista sekä lisää liiketoiminnallista aloitteellisuutta,

ovat sopineet seuraavasta:

1 artikla

Määritelmät

Tässä sopimuksessa

1. "Sijoitus" tarkoittaa kaikenlaista varallisuutta, jonka sopimusvaltion sijoittaja on luonut tai hankkinut toisen sopimusvaltion (isäntävaltion) alueella isäntävaltion lakien ja määräysten mukaisesti, ja siihen kuuluu erityisesti joskaan ei yksinomaan:

a) irtain ja kiinteä omaisuus sekä omistusoikeudet kuten kiinnitykset, panti- ja pidätysoikeudet sekä leasing-sopimuksella vuokrattu omaisuus,

AGREEMENT

BETWEEN THE GOVERNMENT OF THE REPUBLIC OF FINLAND AND THE GOVERNMENT OF THE REPUBLIC OF ALBANIA ON THE PROMOTION AND PROTECTION OF INVESTMENTS

The Government of the Republic of Finland and the Government of the Republic of Albania, hereinafter referred to as the Contracting Parties;

desiring to intensify economic cooperation to the mutual benefit of both countries and to create fair and equitable conditions for investments by investors of one Contracting Party in the territory of the other Contracting Party,

recognizing that the promotion and protection of investments favour the expansion of economic relations between the two Contracting Parties and stimulate business initiatives,

have agreed as follows:

Article 1

Definitions

For the purposes of this Agreement;

(1) The term "Investment" means any kind of asset, established or acquired by an investor of one Contracting Party in the territory of the other Contracting Party (host Party) in accordance with the laws and regulations of the host Party, and includes particularly, though not exclusively:

(a) movable and immovable property as well as property rights such as mortgages, liens, pledges and leases;

Figure 1. Examples of sentence-level (left) and column-level (right) multi-lingual documents

3. XML and Automatic Text Structure Recognition

Once we had all the sources in machine-readable format, we augmented them with mark-ups for (1) layout, (2) text structures, and (3) linguistic information.

Even though our source data comes in a variety of formats, the layout of various texts possesses a strong commonality; they contain text blocks and paragraphs. Therefore, we first converted all our multi-format documents into an XML mark-up for their layout structures. This XML layout is language-independent.

Secondly, we enhanced these layout mark-ups to also reflect document structure. International investment agreements comprise preface (such as title page and table of contents), preamble, text body (i.e. article and paragraph), conclusion (i.e. signatures), and some times attachments (e.g. annex). Hence, we divided each document into these five text zones and annotated the zones with XML mark-up.

Listing 1: Example of XML mark-up for document structures

```
<?xml version="1.0" encoding="utf−8"?>
<treaty>
    <main language="en">
        <preface/>
        <preamble><p>The Government of the Republic of Turkey and ...;</p></preamble>
        <body>
            <div type="article" num="1" title="Definitions"><intro><p>For the purpose ...;</p></intro></div>
        </body>
        <conclusion><p>IN WITNESS WHEREOF, ... </p></conclusion>
```

[11] Most of our documents contain less than 3 languages

```
        <attachments/>
    </main>
</treaty>
```

For the segmentation of these five text zones, we compiled a set of corresponding, typical linguistic feature patterns. For instance, a preamble typically begins with "The Government of the [county name]..." and ends with "... have agreed as follows;" or "... to conclude the following agreement;." Signatures begin with "In witness thereof ..." or "Done in duplicate at ..." Our linguistic feature method is similar to those for legal definition extraction [4], norm classification [5], and content zone identification of German court decisions [6].

In the text body and attachments, we automatically structured texts at the level of articles and paragraphs. This segmentation was again based on the surface patterns of enumeration structures, such as [alpha-numeric character(s)] + period, to recognize the beginning of items of enumerations such as *1.* or *1a.*.

Thirdly, we further enriched this XML mark-up with linguistic information for English. For this purpose, we used a tool of natural language processing [7]. Texts were automatically segmented into sentences, then the sentences were tokenized and augmented with part-of-speech tags, named entities, and dependency grammar structures. This linguistic information allows us to extract further information, for example, meta information such as the date of the signatures or legal definitions (cf. [4]).

4. Conclusion

In this paper, we described our approach to constructing a corpus of international investment agreements. We illustrated that language technology can contribute significantly to reducing the manual effort when building a corpus based on multi-lingual and multi-format source data. In our future work, we will develop an information retrieval system that enables investment law scholars, arbitrators, negotiators, and other legal practitioners to easily retrieve investment treaties.

References

[1] Kenneth J. Vandevelde. A brief history of international investment agreements. *U.C.-Davis Journal of International Law & Policy*, 12(1), 2005.

[2] The role of international investment agreements in attracting foreign direct investment to developing countries. UNCTAD series on international investment policies for development, United Nations, New York and Geneva, 2009.

[3] Kenneth Heafield, Rohan Kshirsagar, and Santiago Barona. Language identification and modeling in specialized hardware. In *Proceedings of the ACL-IJCNLP 2015*, 2015.

[4] Stephan Walter. Linguistic description and automatic extraction of definitions from German court decisions. In *Proceedings of the LREC*, 2008.

[5] Emile de Maat and Radboud Winkels. A next step towards automated modelling of sources of law. In *Proceedings of the ICAIL-09*, 2009.

[6] Florian Kuhn. A description language for content zones of German court decisions. In *The proceedings of Workshop SPLeT-2010*, 2008.

[7] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the ACL: System Demonstrations*, 2014.