


Automatic Labeling of Articles in International Investment Agreements

Machine Learning and Word Embeddings

Conference Poster

Author(s):

Rao, Susie Xi 

Publication date:

2017-12-14

Permanent link:

<https://doi.org/10.3929/ethz-b-000228390>

Rights / license:

In Copyright - Non-Commercial Use Permitted

Automatically Labeling Articles in International Investment Agreements

Machine Learning with Word Embeddings



Susie Xi RAO

ETH Zurich, KOF Swiss Economic Institute
University of Zurich, Institute of Computational Linguistics
rao@kof.ethz.ch

1 Abstract

The author investigates various **machine learning (ML)** methods using features such as **bag-of-words (BoW)** and **word embeddings (EMB)**. Comparing the performances of supervised and unsupervised systems leads to the conclusion that word embeddings can effectively expand the **semantic features for words and documents**, which enables the accurate categorization of texts from closely related sub-fields of one research area. To the best of the author's knowledge, this work is the **first** endeavor to categorize treaty articles using the whole body of IIA text.

2 Highlights

- International investment agreements (IIAs)** are international commitments amongst contracting parties to protect and promote investment. Although each treaty has a distinctive structure regarding placement and organization of information, IIAs as instruments of international law share underlying textual and legal structures. Treaty articles are important components in IIAs: **Some articles have been assigned with titles, while the other articles remained untitled**. More details on corpus see [1]
- Pre-trained word embeddings** from *GoogleNews* were used as features. The usage of pre-defined textual categories (see Figure 5e) and their definitions as "centroids" in clustering is effective. Various machine learning methods were tested and compared.
- It is believed that treaty article categorization can assist mapping treaty texts to their inherent structures. The resulting simplified structure of a treaty is represented by IIA topics, which is beneficial to organizing treaties in information retrieval systems or databases.

3 Corpus & Pipeline

Figure 1: Possible structures in XML where articles are stored

```
<div type="part" num="1" title="GENERAL PROVISIONS">
<div type="title" num="1" title="OBJECTIVES, PRINCIPLES AND ACTORS">
<div type="chapter" num="1" title="Objectives and principles">
<div type="article" num="1" title="Objectives of the partnership">
<p>... Systematic account shall be ...</p></div>
<div type="article" num="2" title="Fundamental principles">
<p>L 317/7</p></div></div></div></div></div>
```

Figure 2: Lowercased token and type counts for titled and untitled articles

	article	token	type
untitled	10,074	1,809,743	22,304
titled	34,524	7,505,258	38,953
total	44,598	9,315,001	52,171

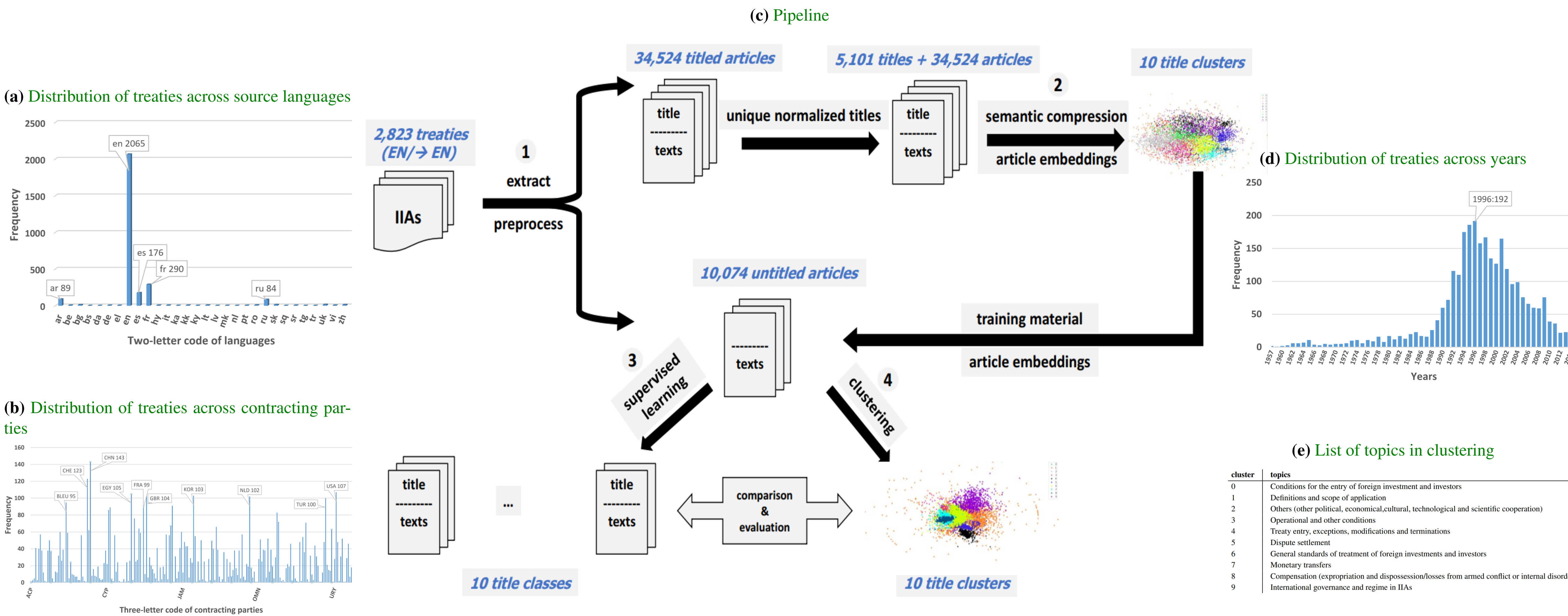
Figure 3: Example of a titled article

ARTICLE 1
Definitions
For the purposes of this Agreement:
1. "Investment" means every kind of asset that is invested by investors of either Party in the territory of the other Party, in accordance with laws and provisions of this Party.
In particular, though not exclusively, the term "Investment" means:

Figure 4: Example of an untitled article

Article 1
The Parties shall, subject to the laws and regulations and investment policies in force in their respective countries, take all appropriate measures to facilitate, promote, strengthen and diversify trade, economic relations and investment, with the aim of achieving a mutually beneficial expansion of trade, economic relations and investment.

Figure 5: Basic statistics and processing pipeline



4 Results and Error Analysis

Figure 7: The Overall accuracies

Features	ML techniques						
	KNN	linear SVM	non-linear SVM	MLP	SGD	CNN	semi-k-means
BoW	4%	10%	15%	2%	11%	/	/
randEMB	/	/	/	/	/	46%	/
pretrainEMB	57%	53%	58%	52%	51%	/*	/
retrainEMB	18%	17%	18%	19%	17%	/*	60%

/ nonavailable * Till the time of the poster, run 2 epochs, acc: ~20%

Figure 8: Comparison of the precision, recall and f-score

class	non-linear SVM (BoW)			non-linear SVM (pretrainEMB)			CNN (randEMB)			semi-k-means (retrainEMB)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
1	13%	39%	20%	0%	0%	0%	0%	0%	0%	88%	39%	54%
2	14%	10%	12%	58%	70%	64%	100%	10%	18%	50%	70%	58%
3	0%	0%	0%	10%	11%	11%	0%	0%	0%	21%	33%	26%
4	0%	0%	0%	40%	67%	50%	67%	67%	83%	83%	83%	83%
5	100%	13%	22%	73%	100%	84%	73%	100%	84%	100%	88%	93%
6	50%	22%	30%	68%	83%	75%	44%	87%	59%	100%	43%	61%
7	0%	0%	0%	71%	100%	83%	83%	100%	91%	100%	60%	75%
8	0%	0%	0%	78%	100%	88%	78%	100%	88%	78%	100%	88%
9	25%	8%	12%	100%	54%	70%	100%	8%	14%	55%	85%	67%
avg	20%	9%	10%	50%	58%	52%	54%	47%	42%	68%	60%	60%

Figure 9: Error analysis of three samples from class 5

source format	source lang	text	gold label	KNN	linear SVM	BoW non-linear SVM	MLP	SGD	pretrainEMB linear SVM	non-linear SVM	MLP	SGD	CNN	retrainEMB	semi-k-means
msWord	bg	Each Contracting Party hereby consents to submit any legal dispute arising between that Contracting Party and a national of the other Contracting Party concerning an investment of that national ... in case a Contracting Party is not a Contracting State to the said Convention...		5	5	1	1	5	5	5	5	5	5	5	3
html	en	1. Any dispute between an investor of one Contracting Party and the other Contracting Party in connection with an investments shall ... 2. If a dispute between one Contracting party and an investor of the other Contracting Party cannot amicably be settled within six (6) months, ... 3. An international arbitral tribunal shall be constituted ... 4. Each party to the dispute shall bear the cost.	5 Dispute settlement	0	4	1	3	3	5	5	5	5	5	5	
html	en	1. Any investment dispute shall be notified by the investor of one Contracting Party ... 2. In the absence of an agreement between the parties to the dispute within six months of the date of its notification, the dispute shall ... 3. A Contracting Party which is a party to a dispute shall not, ...		0	4	0	4	4	5	5	5	5	5	5	

5 Conclusions

Systems with word embeddings as features outperform those with BoW features. Semi-supervised system outperforms the supervised systems and is easier to train: the best supervised system (non-linear SVM) with pretrained word embeddings could not outperform the semi-supervised k-means clustering with retrained word embeddings tailored to the IIA corpus.

6 Future Work

We can train semi-supervised clustering of the titled and untitled parts in the same model with word embeddings as features and try other unsupervised methods, e.g. affinity propagation, hierarchical clustering.

Acknowledgment

Susie Xi Rao is a recent graduate from the master programs in Multilingual Text Analysis and Computational Linguistics with the Institute of Computational Linguistics, the University of Zurich. She currently works at the KOF Swiss Economic Institute, ETH

Zurich. The author appreciates the valuable comments and support from Prof. Dr. Martin Volk and Dr. Kyoko Sugisaki from the Institute of Computational Linguistics at the University of Zurich and Prof. Dr. Peter Egger from KOF Swiss Economic Institute, ETH Zurich, who has been in charge of manually assigning titles to the treaty articles as gold standards in the evaluation. The author also thanks for the support of the SNIS project *Diffusion of International Law* and numerous feedback from the UZH CL community.

References

- [1] Kyoko Sugisaki, Martin Volk, Rodrigo Polanco, Wolfgang Alschner, and Dmitriy Skougarevskiy. Building a corpus of multilingual and multi-format international investment agreements. In *Legal Knowledge and Information Systems - JURIX 2016: The 29th Annual Conference*, pages 203–206, 2016.