GarNLP: A Natural Language Processing Pipeline for Garnishment Documents

Ilaria Bordino, Andrea Ferretti, Francesco Gullo, Stefano Pascolutti

UniCredit, R&D Department

{ilaria.bordino,andrea.ferretti2,francesco.gullo,stefano.pascolutti}@unicredit.eu

- A **garnishment** is a legal procedure by which a **creditor** can collect what a **debtor** owes by requiring to confiscate a debtor's property that is hold by a third party, i.e., a **garnishee**
- GarNLP is a **natural-language-processing** framework to support a garnishee in processing a large-scale flow of garnishment documents
- **Main tasks**
  1. *Document categorization*
  2. *Information extraction*
- **Benefits**
  - *faster and better management* of garnishment notices: 100s per doc
  - better use of *human resources*: -25% manual management
  - reduced *risks and potential losses* from human errors
  - *applicability* to other languages and applications

Scenario

Garnishment at a glance

- A **garnishment** is a drastic measure for collecting a debt
- Common garnishment: amount **confiscation** from a debtor's checking account
- Common garnishees: **banks**, **credit institutions**
- Garnishees are obliged to a **truthful collaboration** with judicial authorities
- Garnishees must block the garnishment amount in a debtor's checking account

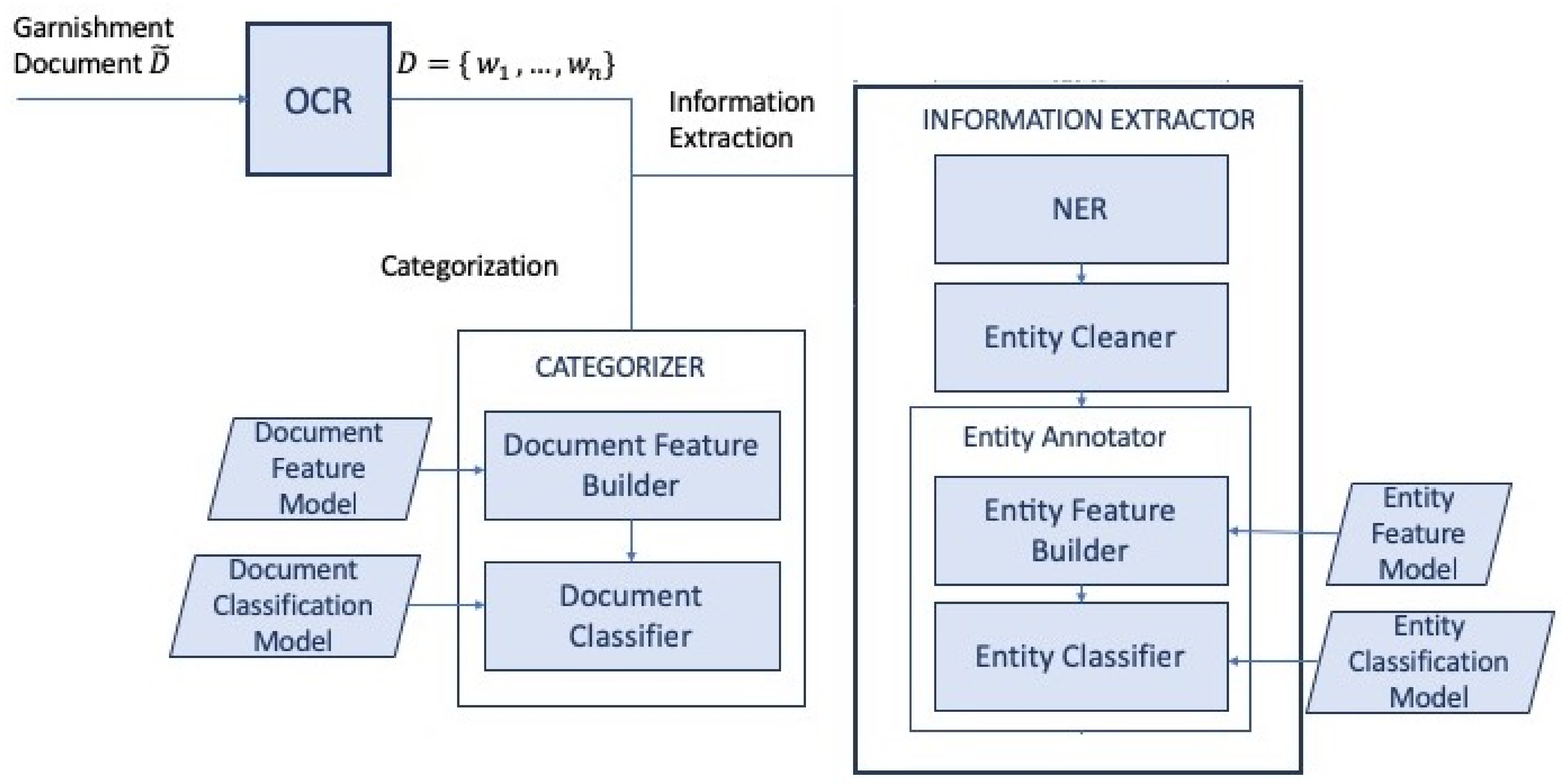
Garnishment from a garnishee perspective

- Upon receiving a garnishment document, a garnishee should perform:
  - **Document processing**: categorizing the document and extracting relevant information
  - **Implementation**: taking the actions requested in the text, e.g., providing information to a legal entity, seizure/release of an account
- We focus on the **document-processing** tasks, i.e., **categorization** and **information extraction**

Main tasks

- **Categorization**: assign a document the correct garnishment-specific category
  - **Practice**: request of information about a debtor
  - **Assignment**: request of seizure of a certain amount
  - **Renunciation**: denial of a previous assignment
- **Information extraction**: extract relevant information from a garnishment document
  - **Actors**: *creditor*, *debtor*, *lawyer*
  - **Amounts**, **Dates**, **Codes**

Framework



- **Categorizer**: *multi-class supervised learning* task
  - **Document Feature Builder**: assign a document D a k-dimensional integer vector $\mathbf{v}(D)$ encoding the *frequency* of the most *discriminant* terms
  - **Document Classifier**: exploit the vectorial representation $\mathbf{v}(\hat{D})$ of every ground-truth document $\hat{D} \in \mathcal{G}$, along with the corresponding $c(\hat{D})$ label, to learn a *document-classification model*
- **Information extractor**: *filter-and-verify*, two-step approach
  - Step 1: *named entity recognition* (NER)
  - Step 2: classify every named entity as an entity of interest or not, while also identifying its type
    - *Entity Feature Builder*: represent the context(s) of an entity via *word embeddings* (in particular, *paragraph vector*)
    - *Entity Classifier*: *multi-class supervised learning*
  - Intermediate *entity cleaning* step: filter out named entities that are easily recognizable as non-interesting
- **GUI**
  - Availability of the original documents (for comparison with information extraction)
  - Possibility to scroll the document up and down to check the information
  - Semi-dynamical highlight to identify the data in che corresponding page
  - Copy/paste enabled

Experiments

Real-world dataset: **101 562** garnishment documents received by the UniCredit bank during **5 months** in **2018**

month	*#docs*	*category*	*frequency*
M1	19 652	*collection agency* (NEQ)	49%
M2	21 827	*private* (NPR)	18%
M3	21 458	*renunciation* (RNC)	16%
M4	17 586	*other* (OTH)	9%
M5	21 039	*assignment* (ASS)	6%
		authority (treasurer) (NET)	1%
		authority (no treasurer) (NEN)	1%

Document categorization on **test set M5**: proposed GarNLP-Cat vs. baselines

training	*method*	*accuracy*
–	HandFirst-Cat	0.727
–	HandFreq-Cat	0.899
M1–M4	HandSup-Cat	0.92
M1–M4	GarNLP-Cat	**0.986**

Document categorization (train M1/M4, test M5): **varying classifiers** in GarNLP-Cat

classifier	*accuracy*
Decision Trees	0.947
Perceptron	0.953
Passive Aggressive	0.959
SGD	0.963
Extra Trees	0.963
Logistic Regression	**0.986**

Actor identification: **classification** results of proposed GarNLP-IE and Freq-IE baseline. Training/test: M1–M4, 80/20, 10-fold

method	*accuracy*
Freq-IE	0.809
GarNLP-IE	**0.926**

Actor identification: performance of GarNLP-IE with **varying the NER model** (Train: M1/M4, Test: M5)

	NER: Tint		NER: GarNLP	
actor	Precision	Recall	Precision	Recall
creditor	0.83	0.301	0.763	0.322
debtor	0.748	0.32	0.781	0.426
lawyer	0.783	0.662	0.758	0.672

Actor Identification: **parameter tuning** of the proposed GarNLP-IE method

W_r	N	*batch size*	*epochs*	*J48*	*Naïve Bayes*	*Random Forest*	*Logistic Regr.*
5	50	500	20	0.545	0.569	0.681	0.849
		500	50	0.504	0.388	0.741	0.856
	100	500	20	0.693	0.686	0.772	0.81
10	50	500	20	0.512	0.39	0.749	0.82
		500	50	0.574	0.48	0.733	0.849
	100	500	50	0.646	0.668	0.778	0.91
		500	100	0.735	0.714	0.886	**0.926**
		1000	100	0.604	0.52	0.73	0.883

Identification of **codes**, **dates**, and **amounts**

entity	*accuracy* (%)	*entity*	*accuracy* (%)
authority	99.32	*court*	95.18
amount	91.06	*injunction #*	40.09
RGE	78.32	*hearing date*	82.63