Classification of Contract-Amendment Relationships

Dr. Fuqi SONG
fsong@hyperlex.ai

Senior Data Scientist @ Hyperlex (hyperlex.ai)

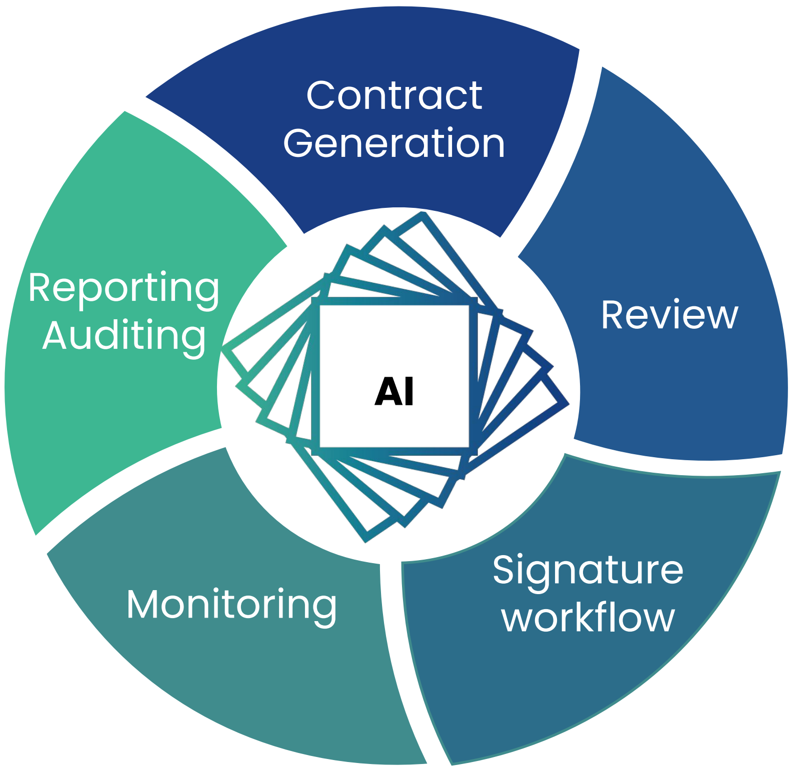
Paris, 25 June 2021

Agenda

- Introduction
- Problem Statement
- Feature Analysis
- Feature Building
- Classification
- Benchmarking and Results
- Applications
- Future Works

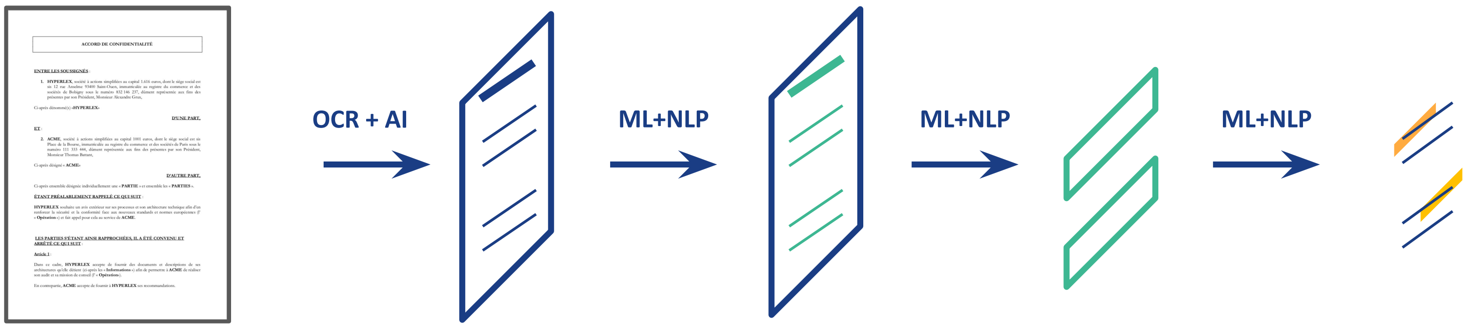
Introduction - hyperlex.ai

- Founded in 2017 based in Paris
- AI-Driven CLM solution providers
  - Contrat generation
  - Review
  - Signature workflow
  - Monitoring
  - Reporting & Auditing



Introduction - hyperlex.ai

hyperlex



Import with different formats (pdf, doc, docx, etc.)

Image processing

Document recognition

e.g. NDA

Clauses recognition

e.g. Clause of responsibilities

Key elements detections

e.g. date of signature, duration of contracts

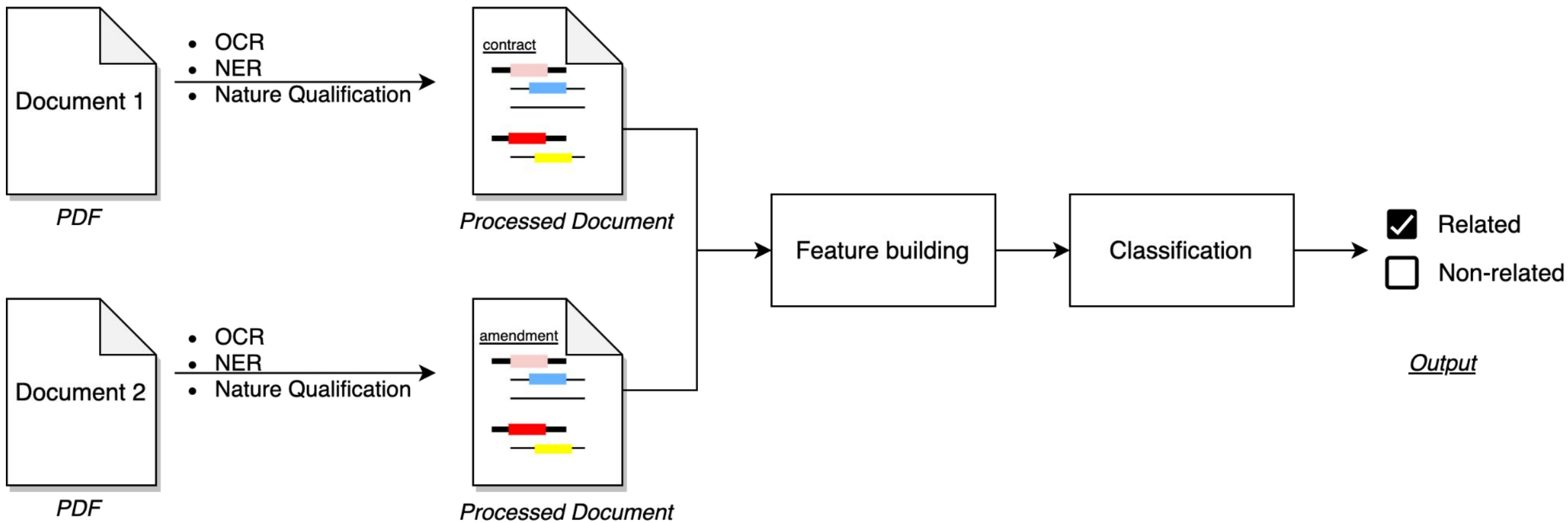
Problem Statement

hyperlex

- Contract-amendment management during the whole life-cycle, for the following typical purposes:
  - when an amendment is added/signed?
  - associated with which master contrat?
  - what terms have been modified?
- An automatic solution is expected to:
  - facilitate the daily jobs of legal practitioners
  - keep track of different due dates and obligations
  - lower legal risks

>> Relationship classification

General schema of relationship classification



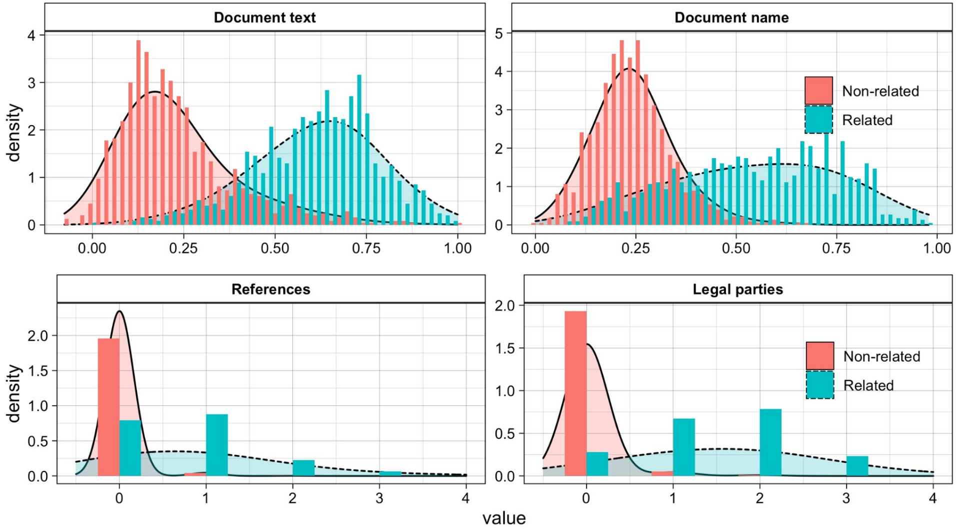
Feature Analysis

hyperlex

- Features allowing to distinguish a pair of related/non-related documents:
  - **Document name**
    - The naming follows certain patterns, e.g. ***Contract No. X12345.pdf*** and ***Contract No. X12345 Amendment 1***.pdf
  - **Document body**
    - The contents are semantically close, e.g. share same contract type and certain clauses
  - **Legal parties**
    - In general, they share the same legal parties
  - **Cross references**
    - Amendments refer certain key information of master contract, e.g. signature date and contract number

Feature Building

- **Document representation**
  - *doc* = *(name, text, legal_parties, references, nature)*
- **Similarity-based feature representation**
  - $\mathbf{f_1}$: document name similarity
  - $\mathbf{f_2}$: document content similarity
  - $\mathbf{f_3}$: number of shared legal parties
  - $\mathbf{f_4}$: number of references



Classification

hyperlex

- **Feature transformation strategy**
  - **Binary:** value -> 0/1
  - **Decimal:** value -> {0.1, 0.2, … 0.9, 1}
  - **None:** no transformation
- **Classifier**
  - Linear SGD (Linear)
  - Random Forest (RF)
  - Multiple Layer Perceptron (MLP)
- **Text embedding**
  - TFIDF
  - FastText

Dataset and baseline

hyperlex

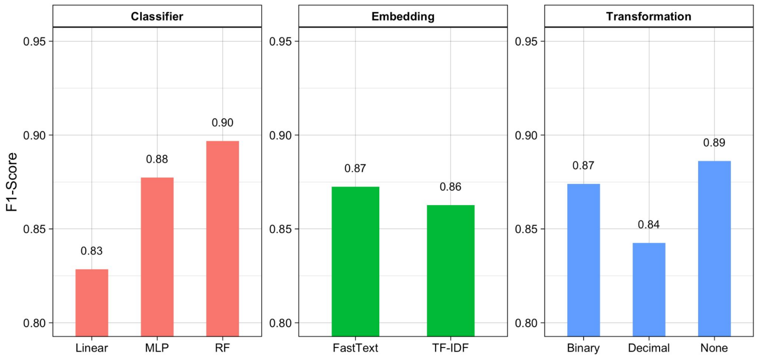
- **Dataset**
  - Annotated by legal experts by showing a pair of possible related documents
  - 1124 pairs of related documents (617 French, 507 English)
  - 1124 paris of randomly sampled non-related documents
- **Baseline** (Empirical and no ML techniques)
  - if document name & text similarity are greater than 0.5
    - related
  - else
    - non-related

Results

- **Baseline**:
  - F1-score: 0.67
- **Best:**
  - F1-score: 0.91
    - **Classifier**: RF
    - **Embedding**: TF-IDF
    - **Feature transformation:** None
- **Other observations**:
  - No significant differences between TFIDF and FastText
  - Classifier RF/MLP works better than Linear SGD
  - Without value transformation works better binary/decimal strategy

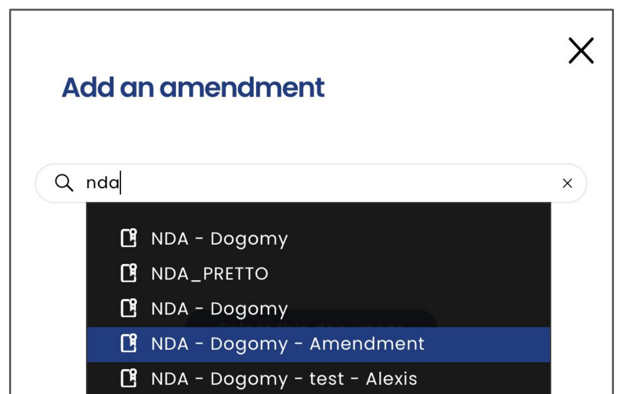
Table 1
Benchmarking results of different configurations on test set

Classifier	Embedding	Transformation	*Precision (%)*	*Recall (%)*	*F1-score (%)*
Baseline	**TF-IDF**	**None**	*77.5*	*64.7*	*67.6*
RF	FastText	Decimal	*90.9*	*87.7*	*89.2*
RF	FastText	Binary	*90.3*	*88.5*	*89.4*
RF	FastText	None	*89.7*	*88.5*	*89.1*
RF	TF-IDF	Decimal	*90.8*	*89.0*	*89.8*
RF	TF-IDF	Binary	*91.3*	*88.3*	*89.7*
RF	**TF-IDF**	**None**	*90.4*	*91.4*	*90.9*
MLP	FastText	Decimal	*89.5*	*85.0*	*87.0*
MLP	FastText	Binary	*89.5*	*87.0*	*88.2*
MLP	FastText	None	*88.8*	*88.6*	*88.7*
MLP	TF-IDF	Decimal	*89.1*	*86.1*	*87.5*
MLP	TF-IDF	Binary	*89.1*	*84.4*	*86.4*
MLP	TF-IDF	None	*89.2*	*88.1*	*88.6*
Linear	FastText	Decimal	*83.2*	*82.5*	*82.9*
Linear	FastText	Binary	*87.9*	*81.9*	*84.4*
Linear	FastText	None	*87.1*	*85.5*	*86.3*
Linear	TF-IDF	Decimal	*84.1*	*65.5*	*69.1*
Linear	TF-IDF	Binary	*86.6*	*86.0*	*86.3*
Linear	TF-IDF	None	*88.0*	*88.2*	*88.1*



Applications in CLM

- Automatic document sorting
  - when the user uploads documents in batch
  - in favor of high precision >> higher probability threshold
- Related documents suggestion
  - when the user uploads a single document
  - in favor of high recall >> lower probability threshold



Future Works

hyperlex

- Reinforce the preprocessing
  - OCR/NER
- Improve the cross-reference detection
  - named linked entity detection
- Explore the textual features
  - adding document content embedding to features
- Fine-tune the parameters
  - similarity threshold
  - train by users

Classification of Contract-Amendment Relationships

Dr. Fuqi SONG
fsong@hyperlex.ai

Senior Data Scientist @ Hyperlex (hyperlex.ai)

Paris, 25 June 2021