First, it is hard to tell the exact solution without the exact definition of a task and the actual data. It is not clear for me how “same” are the regular contracts or how “very different” they are, what is the language of the contracts, etc.? There could be different types of contracts which means, that they will contain different blocks of information and can be structured in the different way.

The stated problem is a problem of data extraction which is often referred as Named Entity Recognition (NER).
NER consists of two substeps:

- Named Entity Identification
- Named Entity Classification

That means first algorithm finds the entities mentioned in a given text and only then it assign them to a particular class in our list of predefined entities.

The language of the legal contacts is usually well organized and companies have their templates, that are modified and used for different contractors. But the good thing is, that structure of the document is usually kept, formats (dates, numbers) are the same and information is presented in the same order. This allows to define and find in document standard “word patterns” to extract information.

Second, you mentioned that the “regular” contract is organized in PDF, but PDF format has different options:

- AcroForms are PDF files that contain form fields
- PDF can contain Text, Tables, checkboxes, pictures
- PDF can be an image stored as in PDF format
- PDF can be password protected
- Structured plain text with blocks like General information, Parties, Object of agreement, Purpose, Definitions, Scope of services, Terms and conditions, Termination, Contacts, Signatures

Different PDF types need different treatment to extract information from.

To illustrate my finding I first break down the problem into meaningful parts and then provide the example of the solution I’ve managed to produce in such a short time period.

In general tasks to solve:

1. Parse the PDF file
2. Locate the part of the document that may contain the needed info – different approaches can be applied to different data fields – it can be needed to find specific info or copy all the block.
3. Locate the sentence that may contain needed info.
4. Locate the needed specific info (dates, names, numbers)

PARSING PDF

The text from PDF can be parsed with Python by a number of libraries

- **http://pybrary.net/pyPdf/**
- **http://www.swftools.org/gfx_tutorial.html**

- **http://blog.didierstevens.com/programs/pdf-tools/**
- **http://www.unixuser.org/~euske/python/pdfminer/**
- **https://textract.readthedocs.io/en/stable/index.html**

Data needed

Training and test samples with texts and the target result

Populating data

1. Data from the contracts that were already processed manually
2. Data from new contracts that are processed
3. Information about the algorithm performance (matches, quality, etc.)

Structure of the data

1. Type of the contract
2. Type of the field to find (dates, company 1, company 2, numbers, terms)
3. Suggestion (can be several)
4. Wrong, right, partly right (degree of match)
5. Correct data from text (if wrong suggestion)
6. Part of the document
7. Full sentence containing right text.
8. Section or paragraph containing the needed info.
9. Language of the contract
10. Technical info - how the fields were filled (manually, algorithm, version of the algorithm, etc.)

Stage One

1. Parse text and select the parts from the document to locate and visualize or the specialist where to take the information from based on the document structure and key words (rule based algorithm).
2. By knowing the structure of the document we can narrow the search (locate) to the specific part of the document.
3. Use regular expressions (regex or regexp) that are extremely useful in extracting information from any text by searching for one or more matches of a specific search pattern

Stage Two

NLP method that can help to better identify

- key words,
- patterns (n-grams),
- part of speech in a patterns,
- pre-trained NER models (spacy, StanfordNER)

Stage Three

Combine two Stages to achieve better results.

For example, if the result by ReGex matches the result from a NER than the lever of certainty is higher.

Extra

1. Make checks if the typed-in information is present in the contract (typo errors, right spelling, etc.)
2. Collect the data about algorithm performance at each step (as mentioned in “Structure of the data”)

RESULTS

Text sample

This agreement is made and entered into by and between 'Abc & Co.' and 'Bcd LLC' for term 1 year starting from April 1, 2020, hereinafter collectively referred to as the Parties. Samuel L. Jackson in the place (New York) and on the date written below, with the following terms and conditions.

Goal: Find date and Companies

Solution 1. Regex (manually defined regex patterns)

April 1, 2020
Company 1: 'Abc & Co.'
Company 2: 'Bcd LLC'

Solution 2. NLTK

ORGANIZATION ----- LLC
ORGANIZATION ----- Parties
PERSON ----- Samuel L. Jackson
GPE ----- New York

Solution 2. NLTK with StanfordNERTagger

ORGANIZATION ---- 'Abc & Co.
ORGANIZATION ---- 'Bcd LLC
PERSON ---- Samuel L. Jackson
LOCATION ---- New York

Solution 4. Spacy

ORG ----- Abc & Co.'
WORK_OF_ART ----- 'Bcd LLC'
DATE ----- 1 year
DATE ----- April 1, 2020
PERSON ----- Samuel L. Jackson
GPE ----- New York

Conclusion

Several approaches were tested. Combining Regex with pre-trained models can provide a solution to assist filling the forms activity.

In the attachment you can find the Python code for the results.

NB! Because the contracts and contract details are the key, and all the NER algorithms make mistakes so I wouldn’t consider the fully-automated option without “eye-control”