



Kairntech Use Case: A Contract Analysis NLP Pipeline

Kairntech webinar, Feb 28, 2023

info@kairntech.com

Introduction

- ❖ *NLP approaches have made huge progress on many use cases in the recent years*
- ❖ *Many previously hard to implement scenarios can now be addressed with impressive quality*
- ❖ *We focus today on how to combine various approaches to implement a contract analysis process*



No code / Low code:
Enable users to make use
of NLP/AI without
requiring them to be
programmers

Adding your expertise

The screenshot displays the Kairntech Sherpa interface with a document analysis view. The document text includes sections like 'ARTICLE 29 - DUREE', 'ARTICLE 30 - LOYER', 'ARTICLE 31 - INDICE', 'ARTICLE 32 - CHARGES', 'ARTICLE 33 - HONORAIRES DE COMMERCIALISATION', 'ARTICLE 34 - DEPOT DE GARANTIE', and 'ARTICLE 35 - DISPOSITIONS PARTICULIERES'. Various values are highlighted with colored boxes and labeled with entity names: '9 ans' (DatedeCommencement), '1er octobre 2014' (DatedeCommencement), '80 958 Euros' (MontantduLoyerannuel), '3 mois' (DatedeCommencement), '2 565 Euros' (MontantduLoyerannuel), and '20 239 Euros' (MontantduLoyerannuel). The right sidebar shows a list of labels with corresponding colored boxes. Three callouts provide instructions: 1. Define your object types (« Entities »), 2. « Show » the system examples of your entities in the documents, and 3. After a while the system starts to generate new ones based on the provided samples.

1 Define your object types (« Entities »)

2 « Show » the system examples of your entities in the documents

3 After a while the system starts to generate new ones based on the provided samples

Scenario today: Contract Analysis

- ❖ Due Diligence, Audit
- ❖ Tedious if done manually: Reviewing numerous complex/lengthy documents
- ❖ Time and cost vs Completeness!
- ❖ So: The scenario is:
 - ❖ We need to analyse Leasing Contracts
 - ❖ But these contracts are buried in a larger collection of other contracts
 - ❖ So we first need to find the Leasing Contracts in the larger set
 - ❖ Then extract key information from these

Our task requires the combination of different approaches:

- 1 Document Categorization: Determine the type of a document
- 2 Entity Extraction: Which specific entities are mentioned in the text?
- 3 Combine individual analysis steps to a larger conditional processing pipeline



Contract Analysis
Pipeline: Categorize,
then extract if
category is right

Import documents

Different unsorted contracts



Manually label documents:
LeaseContract or not?

Train ML models

Train categorization model

Apply model on corpus →
select only LeaseContracts

Final pipeline, implementing the full process

Combine models into Pipeline:
Categorize → right category? →
extract entities

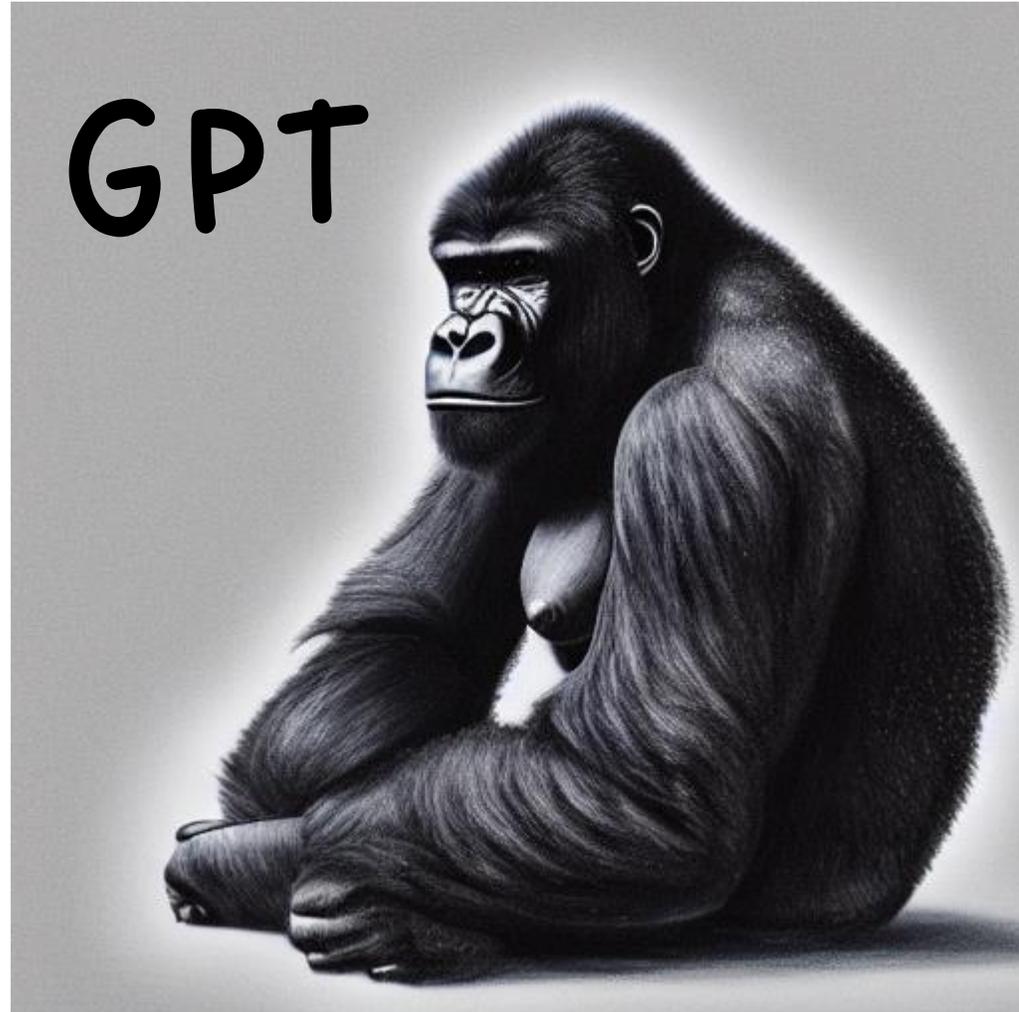
Train entity extraction model

Manually label documents:
Landlord, tenant, date, object?

All inside the system, not requiring any coding

Demo

Whenever you talk about NLP these days:



GPT – needs perhaps no introduction

- ❖ Most popular example of a new class of NLP models (“LLMs – Large Language Models”).
- ❖ Fastest growing piece of software ever: >100mio users after only eight weeks
- ❖ Impressive results on a wide range of use cases: Text generation, question&answer, conversations
- ❖ Proprietary (openAI – Microsoft), others in preparation (Google, Facebook, ...)
- ❖ Apparent potential to disrupt many business models
- ❖ ... potential for our use case here too (Contract Analysis)?

Kairntech as a multi component platform

- ❖ Kairntech integrates wide variety of components and models
 - ❖ DeepL, Spacy, Flair, BERT ...
- ❖ ... allowing to combine them into custom NLP pipelines
- ❖ And of course we have done the same with GPT

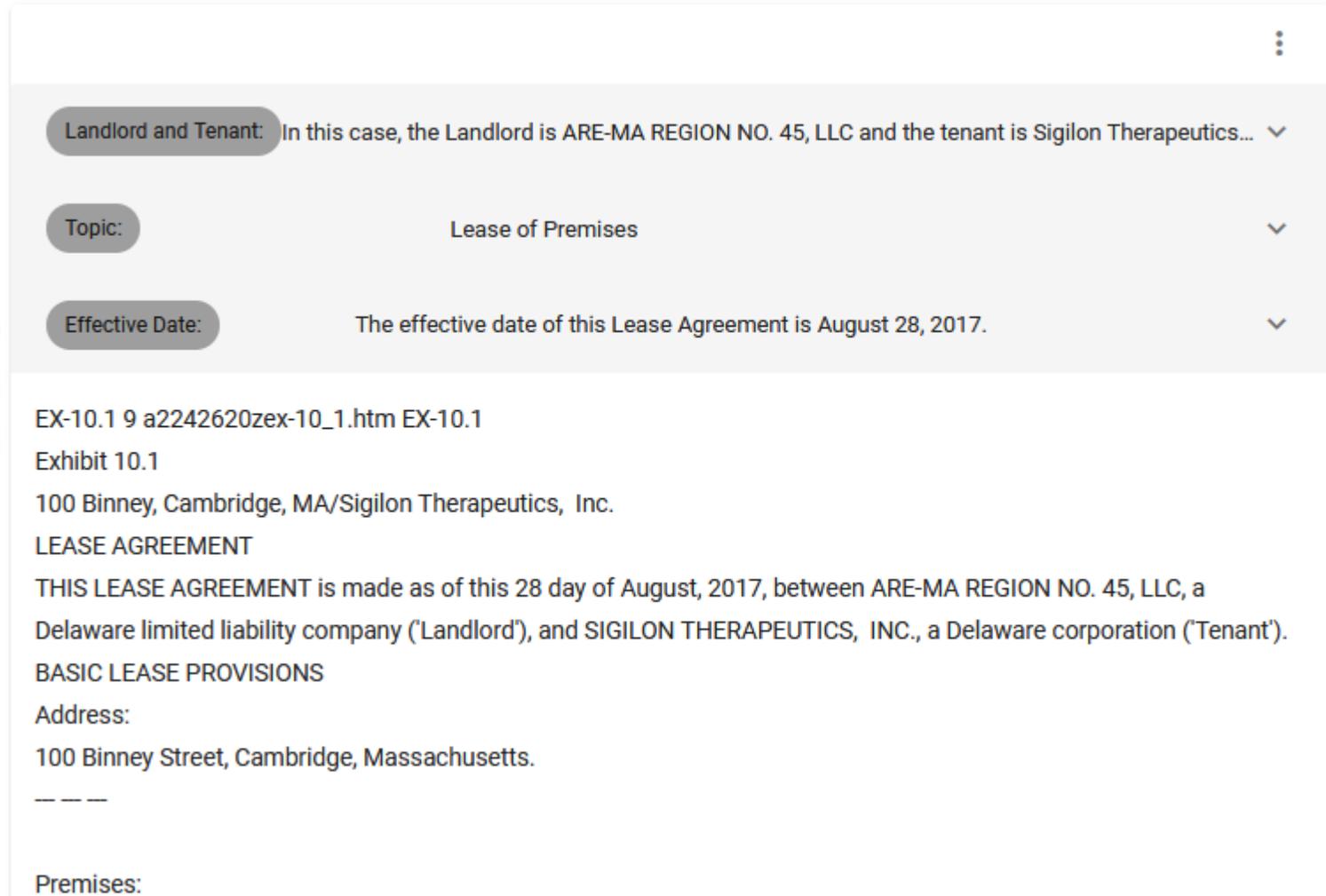
The screenshot displays the Kairntech interface for editing a pipeline named "GPT - Contract analysis". The interface is divided into several sections:

- Name:** GPT - Contract analysis
- Document conversion:** A dropdown menu currently set to "Document conversion".
- Processing pipeline:** A list of three components, each labeled "Off-the-shelf annotator openai_completion". Each component has a blue toggle switch, a gear icon for settings, a blue icon for a specific action, and an 'X' icon for removal.
- Parameters of the openai_completion processor:** A section on the right containing:
 - OpenAIModel:** Set to "text-davinci-003". Below it, the text "An enumeration." is visible.
 - Max Tokens:** Set to "256". Below it, a detailed explanation: "The maximum number of tokens to generate in the completion. The token count of your prompt plus max_tokens cannot exceed the model's context length. Most models have a context length of 2048 tokens (except for the newest models, which support 4096)."
 - Temperature:** A slider set to "1.0".

“Prompt”:

- “Who is the **landlord** and
- “What is the **topic** of the
- “What is the **effective date** of the

Sample result:



The screenshot shows a document analysis interface with a light gray background. At the top right, there is a blue circular button with a white left-pointing arrow. Below it, a white box contains three rows of extracted information, each with a gray pill-shaped label on the left and a dropdown arrow on the right:

- Landlord and Tenant:** In this case, the Landlord is ARE-MA REGION NO. 45, LLC and the tenant is Sigilon Therapeutics...
- Topic:** Lease of Premises
- Effective Date:** The effective date of this Lease Agreement is August 28, 2017.

Below this box, the original document text is displayed in a light gray font:

EX-10.1 9 a2242620zex-10_1.htm EX-10.1
Exhibit 10.1
100 Binney, Cambridge, MA/Sigilon Therapeutics, Inc.
LEASE AGREEMENT
THIS LEASE AGREEMENT is made as of this 28 day of August, 2017, between ARE-MA REGION NO. 45, LLC, a Delaware limited liability company ('Landlord'), and SIGILON THERAPEUTICS, INC., a Delaware corporation ('Tenant').
BASIC LEASE PROVISIONS
Address:
100 Binney Street, Cambridge, Massachusetts.

Premises:

“Nice results! So: Was that it? Can we all go home now?”

Perhaps not quite yet!

GPT without doubt has the potential to disrupt many use cases. To consider when thinking about deploying GPT-powered approaches:

❖ Cost?

- ❖ Model “Davinci”: 0,02\$/1000 tokens text.
- ❖ Sample contract (50000 tokens) → 1\$ per GPT analysis
- ❖ Adapting (retraining) GPT-like models on your content / your vocabularies most often prohibitively complex → combine with other approaches.

❖ Confidentiality?

- ❖ Currently GPT models can normally be run only on openAI infrastructure
- ❖ Comparison with newer, smaller (open source?) models?

❖ Speed?

❖ Accuracy?

- ❖ GPT-like models often seem to “hallucinate” (return unjustified answers)
- ❖ Parameter tuning
- ❖ “Prompt design” vs finetuning: When to favor what?

❖ Very dynamic landscape today: new models being published almost on a weekly basis.

To consider(continued)

- ❖ Randomness
 - ❖ Default GPT returns different results on subsequent identical requests – even where no variation makes sense
 - ❖ (There are ways to reduce that: turn down “temperature”)
 - ❖ Example here: Asking GPT for the total yearly rent:

The amount of rent that needs to be paid in this contract is \$1,692,642.08 (76.00 x 22,746).

The amount of rent that needs to be paid in this contract is \$1,764,973.32 (\$76.00 per rentable square foot of the Premises per year multiplied by 22,746 rentable square feet).

The amount of rent that needs to be paid is \$1,761,264.00 per year as determined by the Base Rent of \$76.00 per rentable square foot of the Premises per year ($\$76.00 \times 22,746$ rentable square feet = \$1,761,264.00).

The amount of the rent that needs to be paid in this contract is \$1,729,284.00 per year. This calculation is based on the rentable area of the premises which is 22,746 square feet and the Base Rent provided in the contract which is \$76.00 per rentable square foot.

The amount of the rent that needs to be paid in this contract is \$76.00 per rentable square foot of the Premises per year.

The amount of rent that needs to be paid in this contract is: \$1,701,359.00. This is calculated by multiplying \$76 per rentable square feet by 22,746 rentable square feet.

To consider (continued):

- ❖ The entity types in our example (Landlord, tenant, date, address) are very common, GPT picked them up out of the box
- ❖ Often, Contract analysis is about more detailed and sophisticated concepts
- ❖ Adding these in a GPT-powered scenario often requires finetuning GPT models (with at least 200-300 examples)
- ❖ Use Kairntech for generating this finetuning data set (and using it to finetune GPT, where this makes sense)?



GPT (and comparable models) are major piece of technology, affecting many use cases. A fast succession of leaps can be expected in the months and years to come, further extending capabilities and impact.

It should be seen as a tool, to be assessed with its strengths and limitations in order to determine whether, where and how to best be employed on your use case.

And now?

- ❖ Learn more about Kairntech and what we do
 - ❖ Website with use cases, blog, product description: www.kairntech.com
 - ❖ Product video: <https://kairntech.com/doc/video/>
 - ❖ Documentation, FAQ: www.kairntech.com/doc

- ❖ Interested to try it out yourself? (no costs, no commitment)
 - ❖ Ask for a free trial www.kairntech.com

- ❖ Any suggestion, question, comment?
 - ❖ info@kairntech.com



*Thank you for your
attention!*

info@kairntech.com