ABBYY®

ABBYY® FlexiCapture® 12

Processing unstructured documents using NLP

Contents

Processing unstructured documents using NLP 3

- Extracting information from texts 3
- Installing the NLP module 4
- Creating a new Document Definition........ 4
- Creating NLP models........ 4
    - Creating document fields 4
    - Creating a segmentation NLP model 6
    - Creating an entity extraction NLP model 7
- Training your NLP models........ 8
- Loading an existing NLP model 9
- Using extraction scripts........ 10
- Known limitations 10

Processing unstructured documents using NLP

Natural Language Processing (NLP) is a subfield of artificial intelligence and computational linguistics. NLP is concerned with computer analysis and synthesis of natural languages. One possible practical application of NLP is the extraction of meaningful data from text.

The way a document is processed depends on its structure. For our purposes, we can distinguish three types of documents: structured, semi-structured, and unstructured documents.

- Structured documents contain a set of well defined data fields whose design, number, and placement do not change from one document to another. Examples of structured documents include forms, questionnaires, and applications.
- Semi-structured documents contain a set of data fields whose design, number, and placement can vary significantly from one document to another. They are also sometimes called "flexible documents." One example of semi-structured documents is invoices, where the number of entries and formatting often depends on the issuing company.
- Unstructured documents contain information that is not structured in any way. They also do not contain explicit data fields. Examples of unstructured documents include contracts, letters, and orders.

  For more information about document types, see Types of documents processed using ABBYY FlexiCapture.

NLP technology should be used to process unstructured documents. For example, NLP can be used to extract the following types of data from a contract: reference numbers, names of parties, important dates (signing date, effective date, term, and termination date), contract price, fees, terms of payment, and so on.

To extract information from tables, structured, and semi-structured documents, other methods should be used (for example, FlexiLayouts).

Extracting information from texts

ABBYY software products use **NLP models** to extract information from unstructured texts. An NLP model tells the program which entities should be extracted from a document. When you train an NLP model on sample documents, the subject area of your texts and the appropriate extraction algorithm are determined so that the information you need can be extracted more efficiently. The effort required to create an NLP model depends on the variety of your documents, the context available to the program, and the complexity and amount of the information that you need to extract.

Extracting data from unstructured texts requires a lot of computing power. Larger texts will take longer to analyze.

However, the necessary information can often be found on a certain page or in a certain paragraph of a very large text. The process of finding such useful parts of text is called **segmentation**. This process requires considerably less time and computing resources than entity extraction, so sometimes you may want to **segment** a document before extracting information from it. For more information about identifying useful segments, see Creating a segmentation NLP model.

To process unstructured documents using NLP, complete the following steps:

1. Install the NLP module.
2. Create a Document Definition.
3. Create and train an NLP model.
4. Alternatively, load an existing NLP model into your Document Definition.

Installing the NLP module

The NLP module is not installed by default. To install it, do the following:

1. Install ABBYY FlexiCapture 12 Distributed Release 3 or later.
2. In the ABBYY FlexiCapture installation directory, open the NLP folder and run the following executable file: *ABBYYFlexiCapture12_Release3_NLP_[installation_version_number].exe*
3. Follow the installation wizard instructions.

Creating a new Document Definition

In ABBYY FlexiCapture, NLP technology is used in conjunction with Document Definitions, which contains information about the locations of various document elements. So before you can use NLP, you need to create a Document Definition and choose the appropriate document type (see the Document Definitions section for detailed instructions).

Creating NLP models

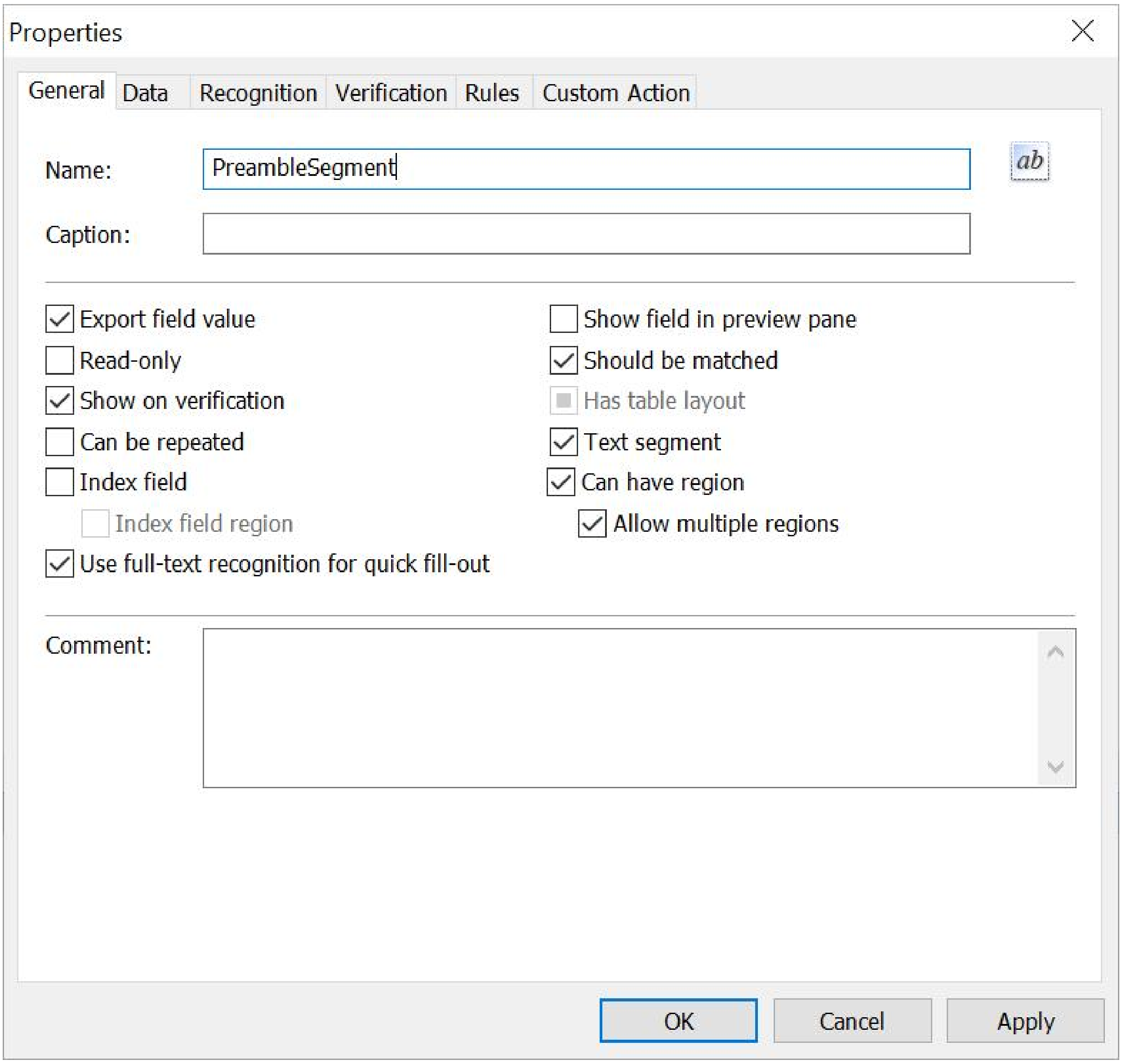
Creating document fields

For every entity that you want to extract, a corresponding field should be created in the Document Definition. To create a field:

1. In the Document Definition Editor, right-click the Document Section name and select **Create Field**.
2. Create a **Text** field.
3. On the **General** tab, select the **Can have region** option.

4. In the **Name** field, specify a name for the field that reflects the nature of the stored data (for example, *PreambleSegment*).

   Important! Field names must not contain spaces or non-English characters or start with a number.



Name: PreambleSegment

Caption:

☑ Export field value
□ Read-only
☑ Show on verification
□ Can be repeated
□ Index field
□ Index field region
☑ Use full-text recognition for quick fill-out

□ Show field in preview pane
☑ Should be matched
■ Has table layout
☑ Text segment
☑ Can have region
☑ Allow multiple regions

Comment:

Repeat the above steps for each entity.

Note: If segmentation is used, a separate text field should be created for each segment.

For each segment from which entities will be extracted:

- Create a non-repeating field in a repeating group.
- Select the **Text segment** option in the field properties.
- Select the **Allow multiple regions** option if some of the segments begin and end on different pages.

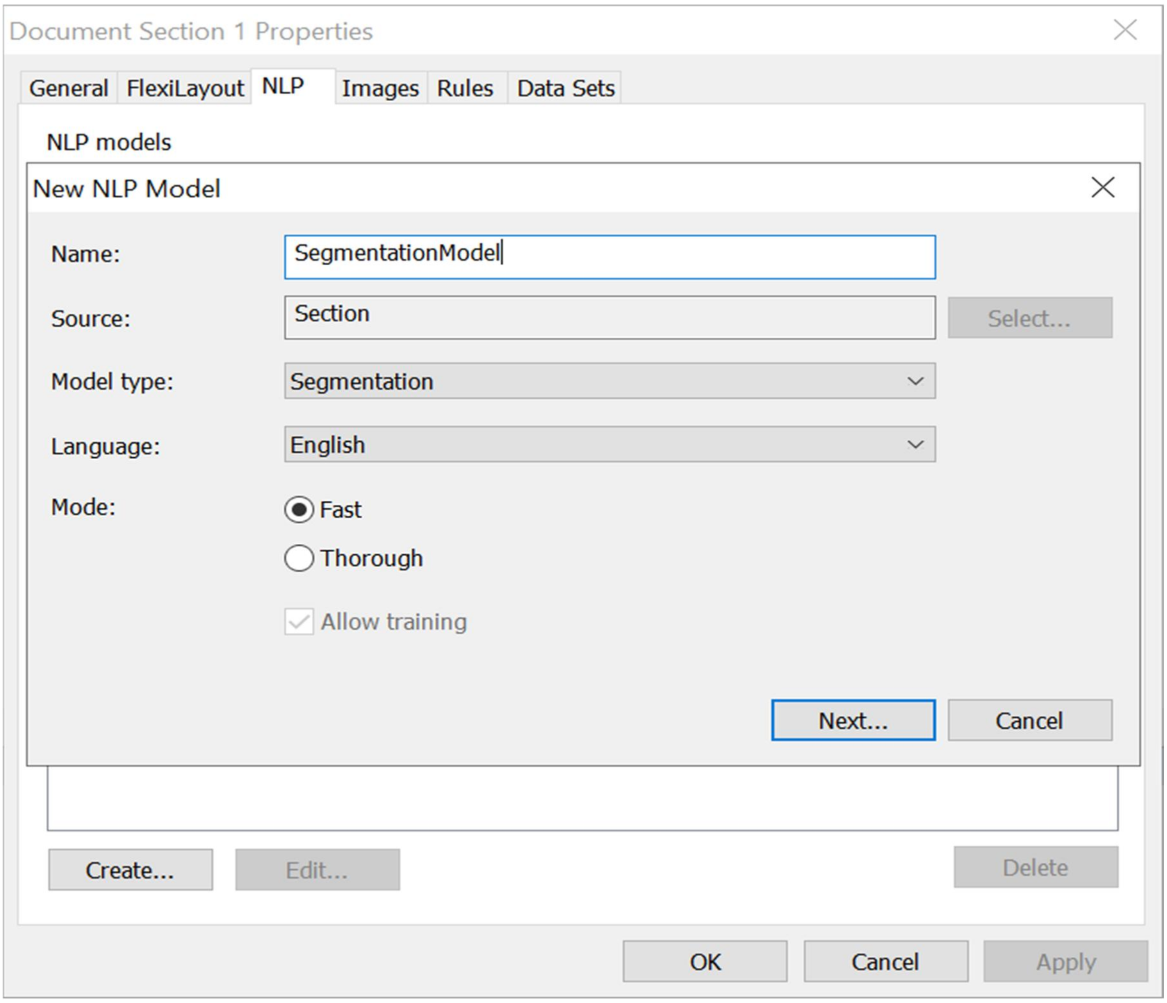
Creating a segmentation NLP model

Segmentation is an optional step, but it improves the accuracy and speed of entity extraction. A special NLP model is required to segment documents.

Important! You can have only one segmentation model for each document section.

To create a segmentation model:

1. In the Document Definition Editor, right-click the Document Section name.
2. Select **Properties...**.
3. In the dialog window the opens, click the **NLP** tab and then click **Create...**.
4. Specify a **Name** for your segmentation model (for example, *SegmentationModel*).
5. In the **Model type** field, choose **Segmentation**.
6. In the **Language** list, select the required language.
7. Select a working mode for your segmentation model. **Fast** mode is recommended for segmentation models.



Document Section 1 Properties

NLP models

New NLP Model

Name: SegmentationModel

Source: Section

Model type: Segmentation

Language: English

Mode: ● Fast ○ Thorough

☑ Allow training

8. Click **Next...**.

9. In the next dialog box, specify all the fields into which the segments will be extracted.

10. Click **OK**.

Once a segmentation model has been created, you need to train it on some sample documents.

Creating an entity extraction NLP model

To extract entities, you need an entity extraction NLP model that has been trained on manually marked up documents. To create an NLP model:

1. In the Document Definition Editor, open the document section properties and click the **NLP** tab.

2. Click **Create...**.

3. Specify a **Name** for your NLP model (for example, *EntitiesExtraction*).

4. As the data source, select either the appropriate section (if no segmentation is used) or the segment (if have chosen to use segmentation).

5. In the **Model type:** field, choose **Extraction**.

6. In the **Language** list, select the required language.

7. Select a working mode for your NLP model. **Thorough** mode is recommended for entity extraction models.

8. Click **Next...**.

9. Choose the result fields that will be extracted from the selected document section or segment.

Repeat steps 1 through 9 for each document segment or section from which entities should be extracted.

10. Click **Document Definition > Save** to save your Document Definition.

11. Click **Document Definition > Close** to close the Document Definition Editor.

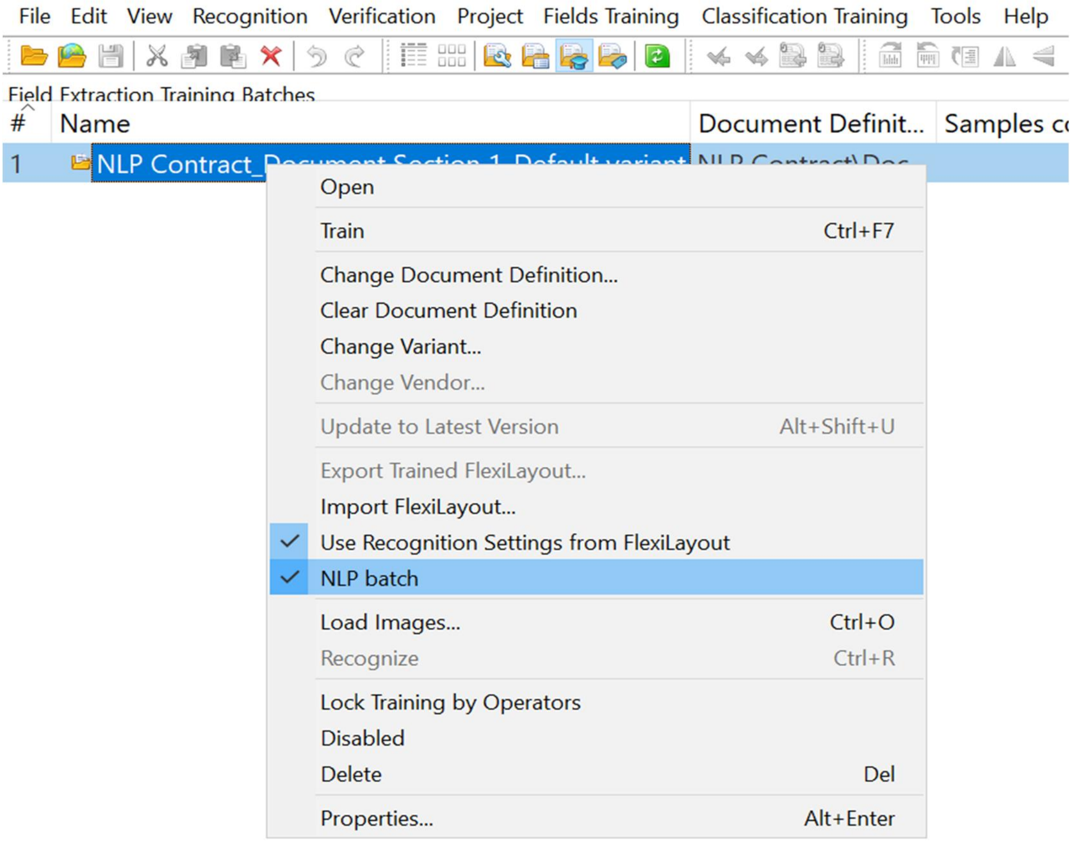
12. Click **Document Definition > Publish** to publish your Document Definition.

Once an entity extraction NLP model has been created, you need to train it on some sample documents.

Training your NLP models

After you have published your Document Definition, close the **Document Definition** dialog box, navigate to the **Field Extraction Training Batches** section and create a new document batch.

1. Click **File** > **New Batch**.
2. In the dialog box that opens, select the Document Definition you have created earlier, then select the section for which you have configured fields and click **OK**.
3. In the **Find Training Batch Variant** window, select the variant to be used for training.
4. Select the newly created batch and either select the **NLP batch** option or click **Field extraction training > NLP batch**.


File Edit View Recognition Verification Project Fields Training Classification Training Tools Help
Field Extraction Training Batches
#
Name
Document Definit...
Samples co
1
NLP Contract_Document Section 1_Default variant
NLP Contract\ Doc
Open
Train Ctrl+F7
Change Document Definition...
Clear Document Definition
Change Variant...
Change Vendor...
Update to Latest Version Alt+Shift+U
Export Trained FlexiLayout...
Import FlexiLayout...
✓ Use Recognition Settings from FlexiLayout
✓ NLP batch
Load Images... Ctrl+O
Recognize Ctrl+R
Lock Training by Operators
Disabled
Delete Del
Properties... Alt+Enter

Now you need to load the documents that will be used to train the NLP model.

1. Open the batch that you have created by double-clicking it.
2. Click **File > Load Images...**.
3. In the dialog box that opens, click **Image Processing Settings...**, select the **One document per file** option, and click **OK**.
4. Choose the documents to be used for training the NLP model.

5. After all the documents have been loaded, select them and click **Recognition > Match Document Definition**. Alternatively, right-click the selection and click **Match Document Definition**. Then choose the appropriate Document Definition.

After you have successfully loaded the documents, you need to manually mark up the fields on each document so that the NLP models will know where to look for entities. To do this, complete the following steps for each document:

1. Double-click a document to open it.
2. Select a field for which information from the document should be extracted. Then either choose the value of the field on the document or draw a rectangle around it. Repeat this step for each field.
3. Go to the next document by clicking the button. Repeat the above steps for all the remaining documents.
4. Save the changes.

After you have marked up all the documents, return to the **Field Extraction Training Batches** view. Right-click the batch and click **Train** on the shortcut menu. Once trained, the model is ready to use.

A trained NLP model can be used in another project. To do this, import the training batch and the corresponding Document Definition into a project of your choice.

Loading an existing NLP model

To load a manually created NLP model into a Document Definition, complete the following steps:

1. Open the Project Setup Station, click **Project > Document Definitions**, choose a Document Definition, and click **Edit...**.

You will see a **Document Structure** pane on the right. If the document structure is not visible, click **View > Document Structure > Show Document Structure**. Alternatively, press **Alt** + **F1**.

2. In the **Document Structure** pane, right-click the section name and select **Properties...** on the shortcut menu.
3. In the **Properties** dialog box, click the **NLP** tab.
4. Load the ZIP file containing your NLP model.
5. Choose the appropriate language.
6. Select a recognition mode. **Thorough** mode is recommended for manually created NLP models.
7. In the document section properties, click **OK**.

Now all fields that can be extracted by the NLP model will be automatically created in the document structure. Set up the data form that will be used by the verification operator (see this section for detailed instructions).

8. Click **Document Definition > Save** to save the Document Definition.
9. Click **Document Definition > Close** to close the Document Definition Editor.
10. Click **Document Definition > Publish** to publish your Document Definition.

Using extraction scripts

Extraction results can sometimes be improved by using extraction scripts alongside the NLP model. You may want to use extraction scripts in the following cases:

- Some of the entities are located in a table while others are located in text fields. In this case, a script will perform better at extracting entities from the table.
- You do not have enough sample documents to train your NLP model.
- Unsatisfactory extraction quality for some of the fields. You may want to use an extraction script for such fields.

Known limitations

Currently, the NLP technology used in ABBYY FlexiCapture has the following limitations:

- Selecting multiple overlapping fields or segments leads to a decrease in extraction quality.
- Selecting a field outside of a segment when the entity corresponding to that field is extracted from that segment will lead to the field not being trained.
- Fields used in training cannot be nested more than one level deep. You can train a field inside a group of fields, but fields inside a group that is part of another group cannot be trained. The nesting level of a field into which an entity should be extracted starts at the source field.
- You cannot create multiple segmentation models within one Document Section
- Multi-lingual NLP models are not supported.