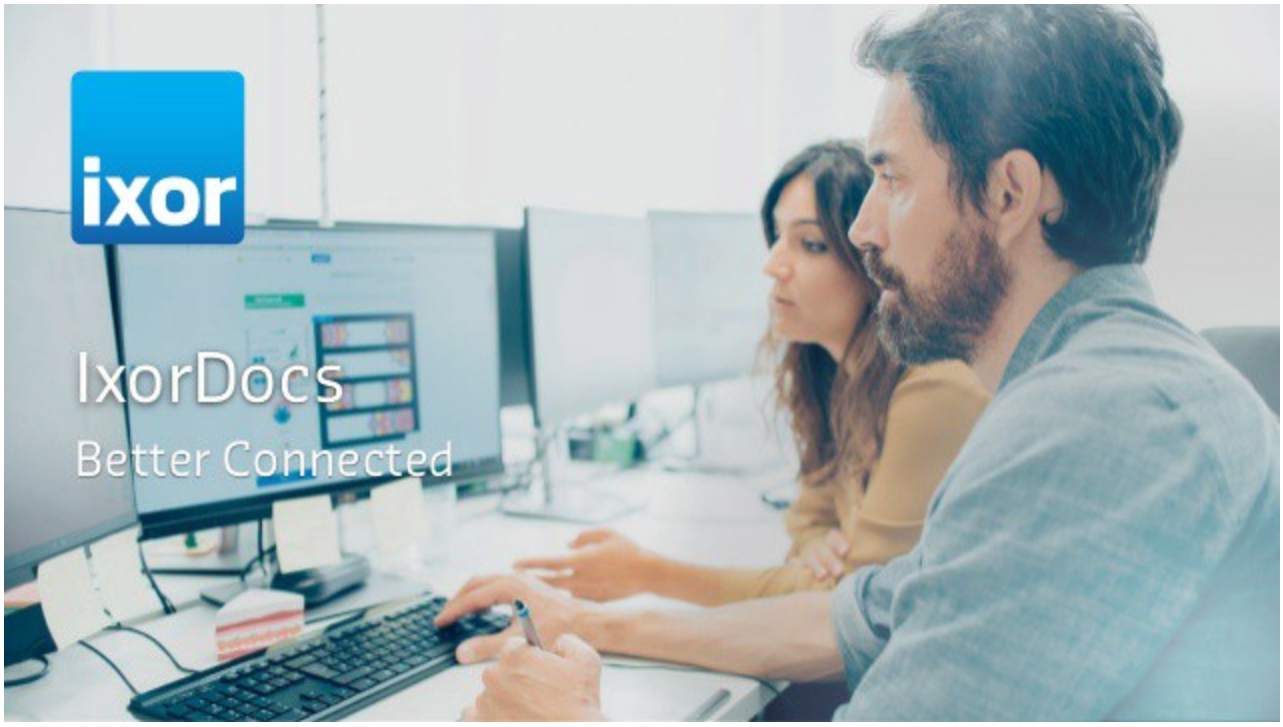
Recognition of Named Entities on Invoices for IxorDocs

Ward Van Laer Follow
Aug 2, 2018 · 4 min read



IxorDocs is the digital link between your company and its clients, employees and the government. Your invoices are sent securely to the correct party, including participants of the international PEPPOL-network. Apart from Business to Government (B2G) and Business to Business (B2B) e-invoicing, IxorDocs can also be used to optimize the HR document flow.

IxorDocs and AI

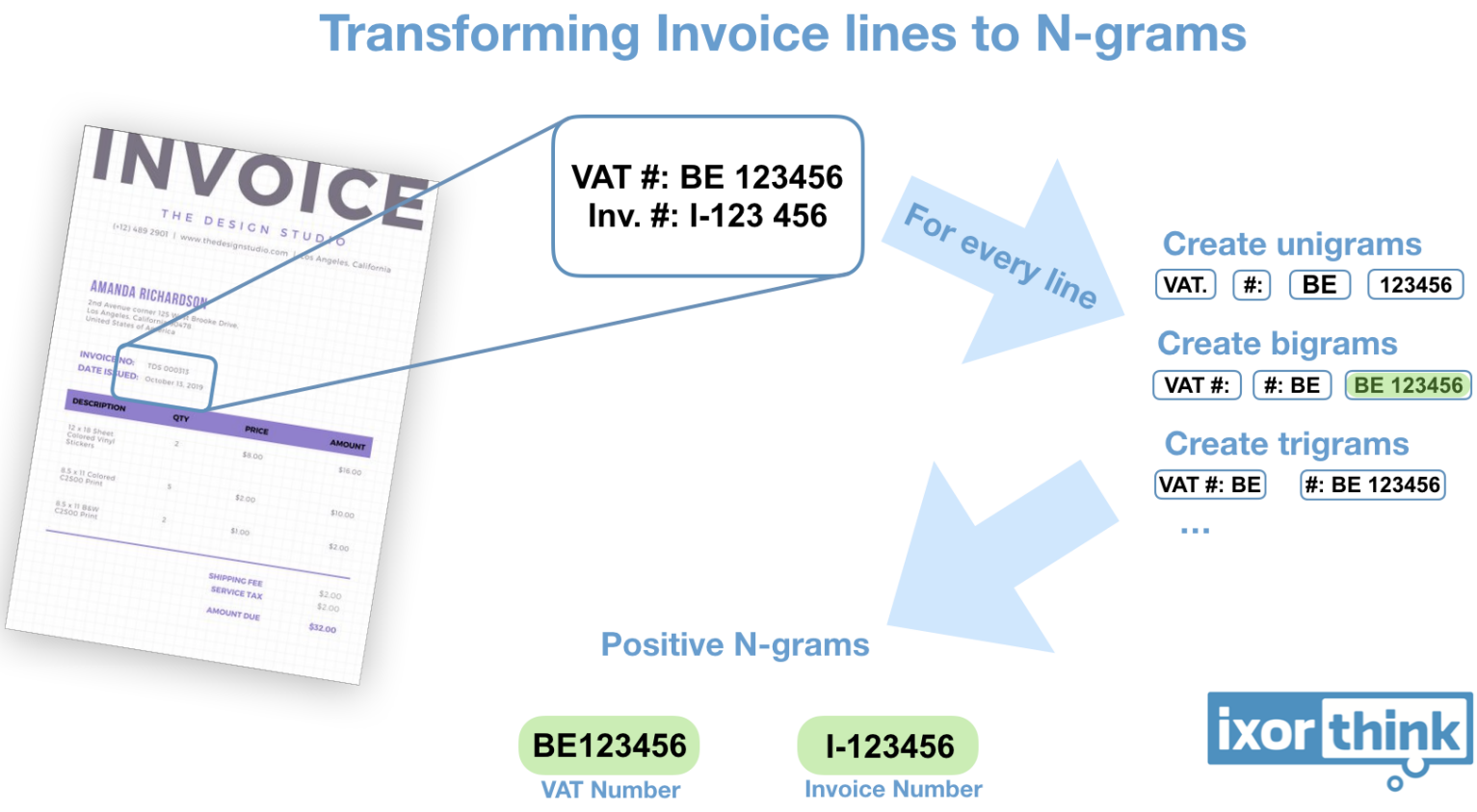
As a part of the document flow we need to transform an invoice in PDF format to UBL, a standard format for digital invoices.
For humans it is not a complex task to recognize a name or VAT number in an invoice; we rely on both the layout, structure and content of the text. However, for a computer this not a trivial task. In order to avoid manual manipulation during the document

processing, AI is the way to go. Using machine learning, it is possible to analyse a PDF invoice in an automatic way. To full auto handle document-flow, all fields of interest, like invoice-number, order-number, date, VAT number have to be detected.

Model overview

Given the small dataset available, it is a challenging task to create a robust model. While it is for most NLP projects a logical choice to go for a connected/recurrent model, it would be impractical to use such model here. An invoice is not constructed of coherent text lines, but mostly contains tables and small text blocks. The location and structure of these blocks will vary over all different invoice templates used. We can already guess that word location and size are important features. This is why we decided to use Decision Tree Classifiers with added context features instead of a recurrent model. Such model will intuitively generate rules like "*the total amount can mostly be found at the bottom of the paper*" if provided with useful features.

To extract all necessary information from the invoices we use the open-source java library PDFBox. Using this Java tool, we convert each PDF file to an array of words including font size and location. Important to notice is that some entities might be a combination of multiple words; e.g. an IBAN number, date or invoice number might contain white spaces. This is why we combine all words to form N-grams before training the classifier. Every created N-gram is afterwards given the correct entity label.



Transforming an invoice to a set of N-grams.

For every N-gram we calculate a feature-vector. Extracted features include word length, font size and position. We also engineer features capturing capitalisation, the amount of number-characters and punctuation. To provide a limited notion of context to the model, we also add previous and following words as features by using the hashing trick. This results in a feature vector of size 164 for every created N-gram, which we will use to classify the N-grams into the entities of interest.

The hashing trick, image by Vivian S. Zhang

Two powerful decision tree classifiers were tested: Random Forest and XGBoost. The Random Forest classifier is known for its simplicity and extremely fast training time. XGBoost goes one step further by using gradient boosting on decision trees and already has a lot of winning Kaggle competitions on its name. On the other hand, XGBoost’s time-complexity scales quadratically with the number of classes (in this case there are 7) which makes it considerably slower than the random forest classifier. For the latter, the amount of classes has no impact on complexity.

To intuitively measure model accuracy we use leave-one-out cross-validation (LOOCV) and the “Recall at k” score, averaged over all invoices in the validation set. For example, “Recall at 1” (R@1) means calculating the percentage of correctly detected entities, when the classifier is given only one shot. XGBoost performs better than expected, but this is at the cost of a ~10x increase in training time.

Scoring results using 4-fold cross-validation and 7 different entities.

When looking at the results per entity, using XGBoost provides the biggest f1-score increase for VAT amount and total. Which are two fields for which the Random Forest model provides only mediocre results.

This is only the start...

With only a small dataset available at the moment, the IxorThink team was able to create and train a named entity recognition (NER) model to correctly analyze new invoices. This can be invoices which follow a known template, or invoices from new customers

based on an unseen template.
At this moment we are able to correctly detect the most important fields in invoices, next we will further roll out this proof-of-concept. We aim to expand the model to detect all invoice lines.
The ultimate goal is to extract useful data from different types of documents, for example credit notes, payslips, etc.

. . .

At IxorThink we are constantly trying to improve our methods to create state-of-the-art solutions. As a software-company we can provide stable and fully developed solutions. Feel free to contact us for more information.

Thanks to Wim De Clercq.

Machine Learning Artificial Intelligence Xgboost Invoicing Software Data Science