Contract Metadata Identification in Czech Scanned Documents

Hien Thi Ha[1], Aleš Horák[1][a] and Minh Tuan Bui[2]
[1]NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic
[2]Le Quy Don Technical University, Vietnam

Keywords: Information Extraction, Scanned Documents, Document Metadata, Contract Metadata Extraction, Czech.

Abstract: Although nowadays digital-born documents are generally prevalent, exchange of business documents often consists in processing their scanned image form as a general human-readable format with one-to-one correspondence to paper documents. Bulk processing of such scanned documents then requires human intervention to extract and enter the main document metadata. In this paper, we present the design and evaluation of a contract processing module in the OCRMiner system. The information extraction process allows to combine layout properties with text analysis as input to a rule-based extraction with confidence score propagation. The first results are evaluated with public Czech contract documents reaching the item extraction accuracy of almost 88%.

1 INTRODUCTION

A contract is a legally binding document that recognizes and governs the rights and duties of the parties to an agreement (Ryan, 2006). Organizations such as companies, institutions, or governmental offices must monitor and handle contracts for a wide range of tasks (Milosevic et al., 2004). Some of them are checking whether obligations, e.g. payments, binding on the party are fulfilled, tracking taxation duties of valuable contracts, or notifying legislation amendments' affects. An important part of such tasks can be automated by extracting contract metadata such as parties involved, dates, or legislation references. However, these pieces of information are mostly filled in management systems manually which is costly and time-consuming.

In a previous work (Ha et al., 2018), the OCRMiner system designed to process scanned invoices based on the combination of layout and text analysis was presented. In the current work, we adapt the system to extract metadata elements from contracts based on a small development set. We also offer an evaluation with detailed analysis of errors.

The next section gives an overview of state-of-the-art in the legal documents processing domain. Section 3 presents a description of the system components with the adaptation to contractual documents. In Section 4, we offer a detailed evaluation of the system with a Czech contract dataset.

2 RELATED WORKS

Research in legal document content classification recently focuses on extracting and classifying clauses, particularly deontic clauses (obligations, prohibitions, and permissions). (Neill et al., 2017) classify deontic clauses using an ensemble of bidirectional long short-term memory networks (BiLSTMs) with the inputs of Google news embeddings. They trained specific legal domain word and phrase embeddings and compared the result with other neural and non-neural classifiers. In a similar task, (Chalkidis et al., 2018) use word embeddings and part-of-speech (POS) tag embeddings trained on an English contract dataset and pre-trained token shape embeddings. The network is also based on BiLSTM but in a hierarchical architecture along with self-attention mechanism to improve training time and accuracy of the classifiers.

In terms of information extraction, (Kwok and Nguyen, 2006) proposed a general template based framework to extract data from PDF contracts. A pattern for each contract data item in a contract type includes data tag, number of words and location (page, paragraph, line, and word numbers). A document type, which is determined by the beginning and ending patterns, identifies a pattern matrix and a list of

[a] https://orcid.org/0000-0001-6348-109X

Ha, H., Horák, A. and Bui, M.
Contract Metadata Identification in Czech Scanned Documents.
DOI: 10.5220/0010243807950802
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 2*, pages 795-802
ISBN: 978-989-758-484-8

contract data tags. There is no specific example of either a contract data pattern or a document type pattern to illustrate the idea in the paper.

In (Winter and Rinderle-Ma, 2018) and (Dragoni et al., 2016), natural language processing (NLP) techniques are used to detect constraints and their relations, or rules in legal documents. In the former, constraints are detected by modal verbs (*shall*, *should*, *must*). These constraints are grouped by either term frequencies or related subjects based on sentence structure or external information. In each group, similarity between each pair of constraints is counted to detect redundant, subsumed, and conflicting constraints using cosine distance of the each constraint word vectors. (Dragoni et al., 2016) use NLP tools to extract rules from legal text. First, they identify deontic components (prohibition, permission, obligation) using a deontic lightweight ontology. Then, these components are combined to create rules using a pattern based model.

The most related works are (Chalkidis et al., 2017; Chalkidis and Androutsopoulos, 2017). In these works, the authors resolve the extraction of contract elements such as contract title, clause headings, parties, dates, values, or legislation references as a sequence labeling task, similar to e.g. named entity recognition (NER). Each sliding window classifier is used for an element type to classify each token of pre-defined extraction zones as positive if it is a part of a contract element and negative otherwise. In the former work, they use Logistic Regression, or Linear Support Vector Machines (SVMs) models. The features involve word embeddings and POS tag embeddings, both pre-trained on a contract dataset, plus hand-crafted features. With the same approach, but using BiLSTM-based models instead of linear ones and with the hand-crafted features being replaced by token-shape embeddings, the latter work improves the previous result. Their best macro average F1 score is 0.88 using a relaxed match. For the contract parties, only the organization name is extracted. The extraction zones, which is up to 20 tokens before and after specified keyword, are explicitly marked in each training and test contract. The system also needs a large amount of data to be annotated for training models.

3 METHODOLOGY

The OCRMiner system pipeline is illustrated in Figure 1. Modules specific for invoice analysis and information extraction were introduced and evaluated in (Ha et al., 2018). Each piece of information is extracted based on a weighted combination of layout and text analysis. The text analysis involves a series of annotations to detect keywords and data types based on either patterns or learning models. Firstly, the contract image is recognized by an OCR tool[1] to obtain words and word positions (bounding boxes). Then, the physical layout including hierarchical elements (lines and blocks) of the pages, block positions in the page and relative positions with neighboring blocks, is built by the *layout analysis* module.

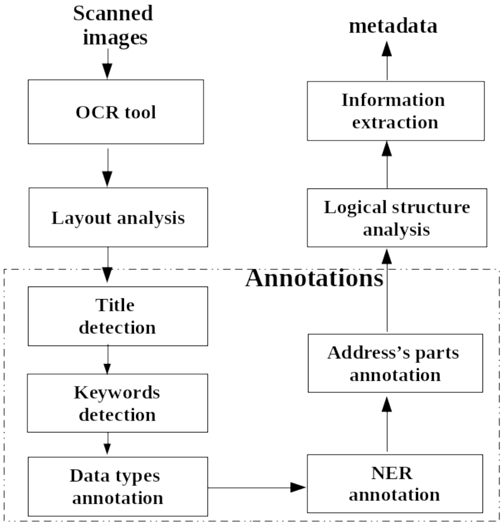


Figure 1: The processing pipeline.

From this point, annotations are added by annotation modules. They involve title, keywords, structural data types, named entities, and parts of addresses. For example, characteristics of the title text are detected by biggest font size, usually center alignment, and containing keyword '*contract*' ('*smlouva*' or its variants in Czech). The first two features are based on layout attributes. For the last one, the text lines are parsed to obtain words and their index forms (lemmata) before searching for the title keyword. Each characteristic increases the confidence score of the item detection. Finally, the candidate with the highest confidence score is marked as the title. If there are more than one with same confidence score, then the first candidate in the reading order, i.e. the one closest to the top of the page, is selected.

Keyword annotation looks for markers of desired data, for example '*contract number*' ('*smlouva číslo*'), '*date*' ('*dne*'), '*address*' ('*se sídlem*'), etc. The list of keywords is prepared based on the most frequent words and word bigrams of the contract dataset adapted using the development set. The key-

[1] The open source OCR system Tesseract (Smith, 2007; Smith et al., 2020) is currently used in OCRMiner.

word search takes into account possible small OCR errors, i.e. it allows a flexible *similarity matching* (see (Ha, 2019) for details). The *data annotation* module searches for structural data types such as a date, VAT number, or legislation reference using regular expressions. In each contract, entity mentions (e.g. an organization (ORG), a person (PER), or a location (LOC)) play an important role, especially in contract party detection. OCRMiner currently uses named entity recognition module based on the Slavic BERT model for 4 languages (Bulgarian, Czech, Polish, and Russian) (Arkhipov et al., 2019), which extends the multilingual BERT model by adding a CRF layer tuned for Slavic languages using Wikipedia and news articles. To improve address recognition, an extra module based on a global address parser Libpostal (Barrentine et al., 2020) is used to detect parts of addresses, such as road/street name, postcode, city, state, or country.

After the annotations, each block is assigned a *block type* in the *logical structure* analysis based on the information gained in the preceding steps using a set of logical rules. These rules are human readable and easy to edit. The reasoning here mimics the human decisions based on visual inspection of the document.

The *information extraction* module concludes the processing to present the identified pieces of information. For each extracted item, the module firstly looks for the item "anchor" in the text, i.e. the corresponding keywords or blocks. Then, in the surroundings of the keyword position, the algorithm searches for the appropriate data type, e.g. a "date" for the invoice date item. The surroundings is limited to either next to the keyword on the same text line, or the text line on the right, or below it. The exact position of the item value is decided by a score weighting function fulfilling the criteria that the block/line contains the data type and does not contain other keywords. Some types of data can be found without keywords such as ORG(anization), PER(son), VAT number, or legislation references. Contract parties are extracted only in blocks being identified as the block type "party", i.e. a block containing at least one keyword in the group of organization, address, contact person, company id, vat number, or bank information, or at least two named entity entries in the corresponding class (PER, ORG, LOC, CITY, COUNTRY, VAT NUMBER). Before parsing a party's information in a block, text blocks that may belong to the same party but that are separated either by physical distance or by covered lines in the block, are joined together using logical rules. The principle here is that if consecutive blocks contain non-overlapping parts of a party's information, then they should be merged together. Each extracted party is assigned a confidence score corresponding to the amount of identified labeled information (ORG, PER, VAT number, company id, or role) in the block.

Table 1: Text statistics of the evaluation contract dataset.

	dev	test	total
№ documents	10	102	112
№ pages	36	589	625
№ blocks	430	8,451	8,881
№ lines	16,587	2,426,298	2,442,885
№ words	147,154	4,911,953	5,059,107

4 EXPERIMENTS

4.1 Dataset

The dataset used for development and evaluation of the contract analysis module of OCRMiner comes from the official state registry of Czech public contracts[2]. The data obtained from the website include contract texts (in PDF) and metadata files (in XML). The registry contains not only contracts but also appendices, price lists, invoices, etc. Therefore, a 2-step filter is applied to select contracts only. The first step automatically filters out documents based on the filename and the text content. The filename usually reflexes the content, so, files having names containing '*obj*' ("objednávka" – order), '*ceník*' or '*cenová nabídka*' (price list), '*příloha*' (appendix) have been removed. Then remaining files have been converted into OCR text. If the text does not contain the keyword '*smlouva*' (contract), then the document is also filtered out. The second step involves manual check. Finally, 112 contracts were selected randomly for the thorough evaluation to be annotated (by one annotator) as the gold standard data. Ten documents are used as a development set and the remaining ones form a test set. Text statistics of the final datasets are enlisted in Table 1.

Although the contracts metadata are available, a further step is still needed to prepare the gold standard data for evaluation. Firstly, the metadata does not contain all the information that is to be extracted such as a representative person or role of a contract party. Secondly, since the registry metadata were entered manually through the available forms, they are in different formats compared to the contract text, especially the dates and addresses. Thirdly, some pieces of information appear in the metadata but not in the

[2] https://smlouvy.gov.cz/

Table 2: Identified items in the contract texts.

Item	№ in dev	№ in test	Example
title	9	102	“Smlouva o poskytování služeb” (*supply of services contract*)
contract type	10	100	“poskytování služeb” (*supply of services*)
legislation	33	547	“$ 1746 a násl. zákona č. 89/2012 Sb., občanský zákoník” (*$ 1746 et seq. Act No. 89/2012 Coll., Civil Code*)
contract number	7	58	”VODA/ZA20-4023”
contract date	8	78	10.1.2020
company name	13	175	“TESCO SW a.s.”
representative	13	164	“Josefem Tesaříkem” (*by Josef Tesařík*)
address	21	194	“tř. Kosmonautů 1288/1, Hodolany, Olomouc, PSČ 779 00”
vat number	10	102	“CZ699000785”
company id	19	191	“25892533”
bank name	6	56	”Česká spořitelna, a.s.”
account number	4	49	”1303699319/0800”
role	19	194	“poskytovatel” (*supplier*)

contract text. For example, contract numbers in some cases are not stated in the original contract but in the metadata only. Moreover, in many contracts, private information is covered, such as an account number or contact details. So, after converting the registry metadata file into the desired format, the data is manually examined before becoming the ground truth for the evaluation.

4.2 Information to Extract

The detected and extracted pieces of the contract information are summarized in Table 2. Specifically, the contract date is the closest date that all parties have signed the contract. Usually, it appears at the end of the contract, before the signatures. If the signature dates are different then the later one is extracted. A contract party is a group of information, involving organization, address, company id, VAT number, a company representative, a party role in the contract, bank name, or an account number. The party role is often stated at the beginning of the party text block, e.g. ‘*zhotovitel*’/*contractor* and ‘*objednatel*’/*customer*, or after the keyword ‘*dále jen*’/*hereinafter*. A full example of information extracted from the first page of a contract is illustrated in Figure 4 in the Appendix.

4.3 Results

Within the evaluation process, each piece of extracted information is evaluated as a *match*, a *partial match*, or a *mismatch*. For all fields except organization and address, the *match* means an exact match. For these two exceptions, a *match* allows to ignore the piece of the gold standard information which is not crucial for the company or address identification.

For example, organization full official name occurrences of:

ground truth:
`Czech Airlines Technics, a.s.`
extracted:
`Czech Airlines Technics`

are considered a *match*.

Differently from the previous use case of invoice information extraction where the parties can always be classified into a seller and a buyer, in contracts the number of parties is not predetermined. Therefore, the extraction process needs to take each text block containing an organization's information as a possible contract party information. In the evaluation phase, each gold standard contract party is compared to each extracted party. The result is recorded for the party having the most common information. This means the evaluation will not search for each piece of individual contract party information in the extracted data as a whole and return a match if such piece is found. The evaluation is here strict in the sense that even if a sought piece of information is extracted but in a different party block then the result will be a mismatch. Some works use a *relaxed match*, i.e. if the extracted information matches the ground truth at a threshold, e.g. 80%, then it is considered as a (relaxed) match. The importance of the missing piece is ignored here. To give an example, if a contract date ground truth is “`1.12.2019`” and the extracted date is “`31.12.2019`”, that makes only 10% difference. However, in the context, the second date was meant as the payment due date, so, it should have been considered as a mismatch instead of a match. Due to such complications, the evaluation is first pre-

Table 3: Test set evaluation results.

Result	№ items	Percentage
Match	1,631	81.14%
Partial match	137	6.82%
Mismatch	242	12.04%
Total	2,010	100.00%

processed automatically using approximate comparisons based on the Levenshtein distance, then examined manually.

The evaluation results of the OCRMiner contract module with the test set are presented in Table 3. Altogether, almost 88% of gold standard information was extracted, with 81% in the exact expected form and approx. 7% with minor differences. Just 12% of items were not identified or identified wrongly.

A detailed evaluation of the individual item types is illustrated in Figure 2. Addresses and contract types have the highest accuracy of 94.3% and 93% respectively. In contrast, contract dates and party roles display the highest number of mismatches with 30.8% and 26.3%. The legislation reference field contains the highest number of minor errors (partial matches) of 15.7%.

In the following section, a detailed error analysis of 50 contracts in the test set identifies and explains the causes of both minor and major mismatches.

4.4 Error Analysis

In the OCRMiner extraction pipeline, the data to extract are identified by keywords, data format or text position. If a keyword is found, then the extraction module looks for the appropriate data item around the keyword based on the visual layout, especially in relation to the keyword position. Non-keyword data are detected by a pattern (e.g. the VAT number) or a pre-trained model (e.g. organization or person name). Therefore, the error causes are classified into different categories: OCR errors, keyword error (there is either no keyword in the text or a new keyword which did not appear in the development set), layout error, named entity recognition (NER) error, block misidentification (extracted in another block), and others. A layout error means the keyword is found but the data text line is not found in the expected relative position, either due to a typing error or the layout match criteria. In the detection of a company name, a keyword is often elided, thus the extraction relies on the NER annotation or the company name's ending. However, the dataset originates in the public sector where many parties are public organizations of a specific area (e.g city, village, etc.). In consequence, NER recognizes only part of the organization name as a location instead of the whole chunk as an organization. For example, '*Město Hostinné*' (Hostinné town) or '*Služby města Náměště nad Oslavou*' (Town services of Náměšť nad Oslavou). As mentioned above, in parties' evaluation the comparison is made for the whole group instead of searching for each piece of information separately causing a mismatch when a piece of information is correctly extracted but assigned to a different block. Furthermore, as we described in 3, the contract title is extracted using 3 criteria involving the font size, the central alignment and a keyword. However, in some cases, the title can be left aligned, or the biggest font is a part of text in the logo or another line. The combination of these errors falls into the title category. Legislation references are identified by flexible patterns consisting of 3 parts: the section mark (§), paragraph ID ('odst. X'), and the act or law ID. But not all 3 parts are obligatory. The act or law is illustrated by number, e.g. 89/2012, or name ('občanský zákoník'/*Civil code*). Full examples are:

Table 4: Error analysis of partial matches.

Error type	items	in %
OCR error	18	39.13
NER	9	19.57
Multi-lines	5	10.87
Pattern	9	19.57
Other	5	10.87
Total	46	100.00

Table 5: Error analysis of mismatches.

Error type	items	in %
In another block	7	6.31
OCR error	31	27.93
Keyword	27	24.32
Layout	10	9.10
NER	12	10.81
Pattern	7	6.31
Title	6	5.40
Other	11	9.91
Total	111	100.00

- `§ 1746 odst. 2 zákona č. 89/2012 Sb., občanský zákoník` (*§ 1746 par. 2 of Act No. 89/2012 Coll., Civil Code*)
- `§ 2586 a násl. zák. č. 89/2012 Sb., občanského zákoníku` (§ 2586 et seq. Act. No. 89/2012 Coll., Civil Code)
- `§ 92a zákona o dani` (§ 92a of the Tax Act)
- `č. 340/2015 Sb.` (No. 340/2015 Coll.)

These patterns are based on findings in the development set with extra flexibility, however, some test set cases yet remained uncovered such as more than one

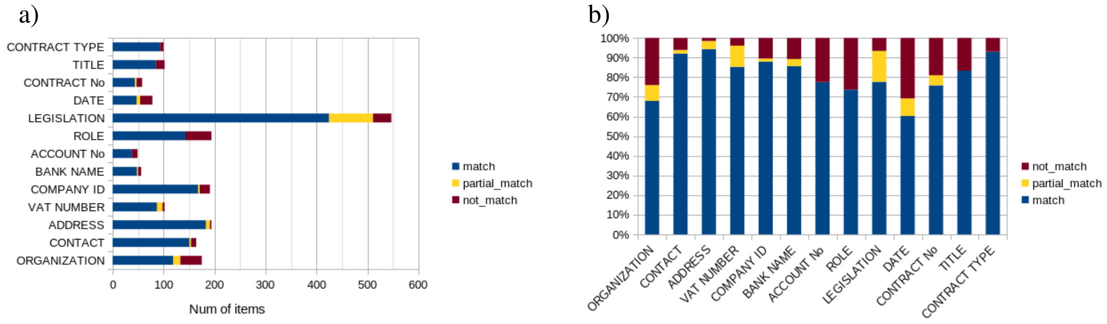
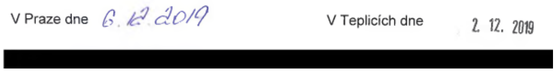


Figure 2: Evaluation of each field: a) by item, and b) by percentage.

section in a legislation reference (§27 a 31 zákona č. 134/2016 Sb.), or a connection word 'násl.'/*seq.* written in the full form ('následujících'/*sequentes*).

The error analysis of each category is summarized in Tables 4 and 5. In the *partial match* section, almost 40% of errors are due to OCR errors, usually in characters sharing similar shapes, e.g. `4`-`A`, `Z`-`7`, or `O`-`0`. The cases where NER and pattern did not detect full organization or full legislation reference caused the same number of errors (19.57%), leaving 10.87% for multiple lines and for other reasons.

In the *mismatch* section, OCR errors caused more than a quarter of mismatches, followed by keyword errors with another 24.32%. As we can see in Figure 2, the accuracy of the contract date item is low. In the analysed contracts, the dates usually appear stamped or hand-written when signing the contract (see Figure 3 for an example). Thus, most of the date errors happen because the OCR engine could not recognize the hand-written characters correctly. In addition, the cover of confidential information sometimes overlapped text in the surrounding areas which made more OCR errors. 10.8% errors was because of NER recognizing a part of an organization name as a location. Layout errors appear in 9% and 6% of items were extracted in a wrong block. The reason was either due to a specific layout design or to distances between information in the group created by a covering black line (as we can see in the example Figure 4 in the Appendix). Pattern errors appear also in 6% followed by title errors (5.41%) and 9.91% of others reasons.


V Praze dne 6. 12 2019   V Teplicích dne 2. 12. 2019

[redacted]

Figure 3: A contract date example.

5 CONCLUSIONS

In the paper, we have presented the first version of the OCRMiner system module for information extraction of scanned contract documents. The design and the architecture of the module have been described in details.

A new dataset for the evaluation of the contract information extraction has been built and used in a thorough evaluation of the contract analysis modules. The evaluation results show that the system is able to identify almost 88% of the contract metadata correctly. A detailed error analysis depicts and classifies the reasons of the current mismatches. Although some modules (e.g. keyword detection) are language dependent, the pipeline is easily adaptable to other languages.

With the presented analysis, the current test set can be seen as an extended development set to evaluate the system on a new and larger test set to confirm its generalization capabilities. This version also offers as a strong baseline for further work where we plan to employ state-of-the-art NLP techniques such as pretrained BERT model tuning on the contract dataset or grammar induction techniques for the layout and content analysis.

ACKNOWLEDGEMENTS

This work has been partly supported by Konica Minolta Business Solution Czech within the OCR Miner project, the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101 and the Specific University Research of Masaryk University.

REFERENCES

Arkhipov, M., Trofimova, M., Kuratov, Y., and Sorokin, A. (2019). Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Barrentine, A. et al. (2020). Libpostal. https://github.com/openvenues/libpostal.

Chalkidis, I. and Androutsopoulos, I. (2017). A deep learning approach to contract element extraction. In *JURIX*, pages 155–164.

Chalkidis, I., Androutsopoulos, I., and Michos, A. (2017). Extracting contract elements. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 19–28.

Chalkidis, I., Androutsopoulos, I., and Michos, A. (2018). Obligation and prohibition extraction using hierarchical RNNs. *arXiv preprint arXiv:1805.03871*.

Dragoni, M., Villata, S., Rizzi, W., and Governatori, G. (2016). Combining NLP approaches for rule extraction from legal documents.

Ha, H. T. (2019). Approximate string matching for detecting keywords in scanned business documents. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019*, pages 49–54.

Ha, H. T., Nevěřilová, Z., Horák, A., et al. (2018). Recognition of OCR Invoice Metadata Block Types. In *Text, Speech, and Dialogue. TSD 2018*, pages 304–312. Springer, Cham.

Kwok, T. and Nguyen, T. (2006). An automatic method to extract data from an electronic contract composed of a number of documents in PDF format. In *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06)*, pages 33–33. IEEE.

Milosevic, Z., Gibson, S., Linington, P. F., Cole, J., and Kulkarni, S. (2004). On design and implementation of a contract monitoring facility. In *Proceedings. First IEEE International Workshop on Electronic Contracting, 2004.*, pages 62–70. IEEE.

Neill, J. O., Buitelaar, P., Robin, C., and Brien, L. O. (2017). Classifying sentential modality in legal language: a use case in financial regulations, acts and directives. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 159–168.

Ryan, F. (2006). *Round Hall nutshells Contract Law.* Thomson Round Hall.

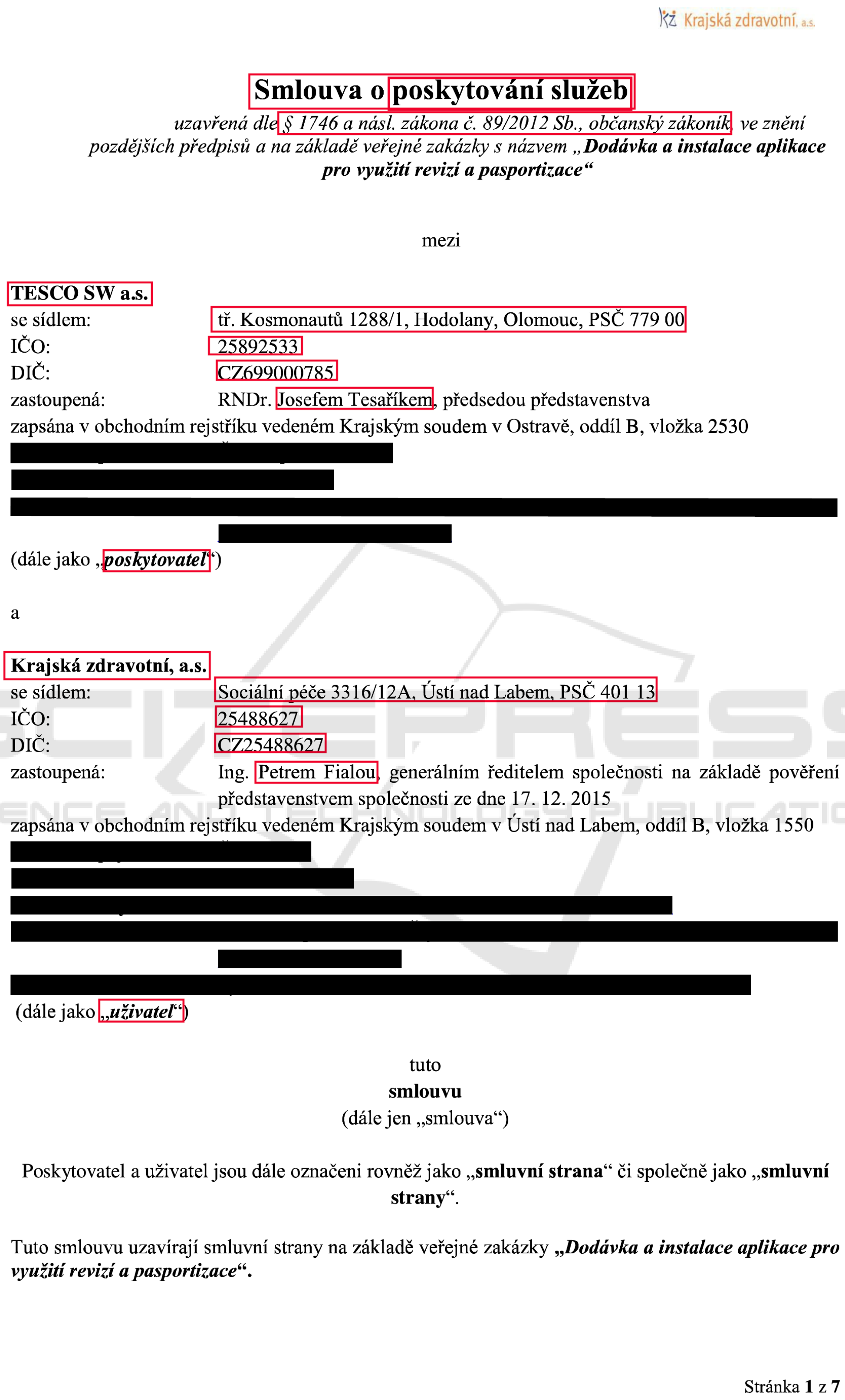
Smith, R. (2007). An overview of the Tesseract OCR engine. In *Document Analysis and Recognition, Ninth International Conference on*, volume 2, pages 629–633. IEEE.

Smith, R. et al. (2020). Tesseract OCR. https://github.com/tesseract-ocr/tesseract.

Winter, K. and Rinderle-Ma, S. (2018). Detecting constraints and their relations from regulatory documents using NLP techniques. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 261–278. Springer.

APPENDIX

An example of a scanned contract with the extracted metadata information is presented on the next page in Figure 4.



Krajská zdravotní, a.s.

Smlouva o poskytování služeb

*uzavřená dle § 1746 a násl. zákona č. 89/2012 Sb., občanský zákoník ve znění pozdějších předpisů a na základě veřejné zakázky s názvem „**Dodávka a instalace aplikace pro využití revizí a pasportizace**“*

mezi

TESCO SW a.s.

se sídlem: tř. Kosmonautů 1288/1, Hodolany, Olomouc, PSČ 779 00
IČO: 25892533
DIČ: CZ699000785
zastoupená: RNDr. Josefem Tesaříkem, předsedou představenstva
zapsána v obchodním rejstříku vedeném Krajským soudem v Ostravě, oddíl B, vložka 2530
[redacted]

(dále jako „***poskytovatel***“)

a

Krajská zdravotní, a.s.

se sídlem: Sociální péče 3316/12A, Ústí nad Labem, PSČ 401 13
IČO: 25488627
DIČ: CZ25488627
zastoupená: Ing. Petrem Fialou, generálním ředitelem společnosti na základě pověření představenstvem společnosti ze dne 17. 12. 2015
zapsána v obchodním rejstříku vedeném Krajským soudem v Ústí nad Labem, oddíl B, vložka 1550
[redacted]

(dále jako „***uživatel***“)

tuto
smlouvu
(dále jen „smlouva“)

Poskytovatel a uživatel jsou dále označeni rovněž jako „**smluvní strana**“ či společně jako „**smluvní strany**“.

Tuto smlouvu uzavírají smluvní strany na základě veřejné zakázky ***„Dodávka a instalace aplikace pro využití revizí a pasportizace“.***

Stránka **1** z **7**

Figure 4: A contract example: extracted information is in the red boxes.