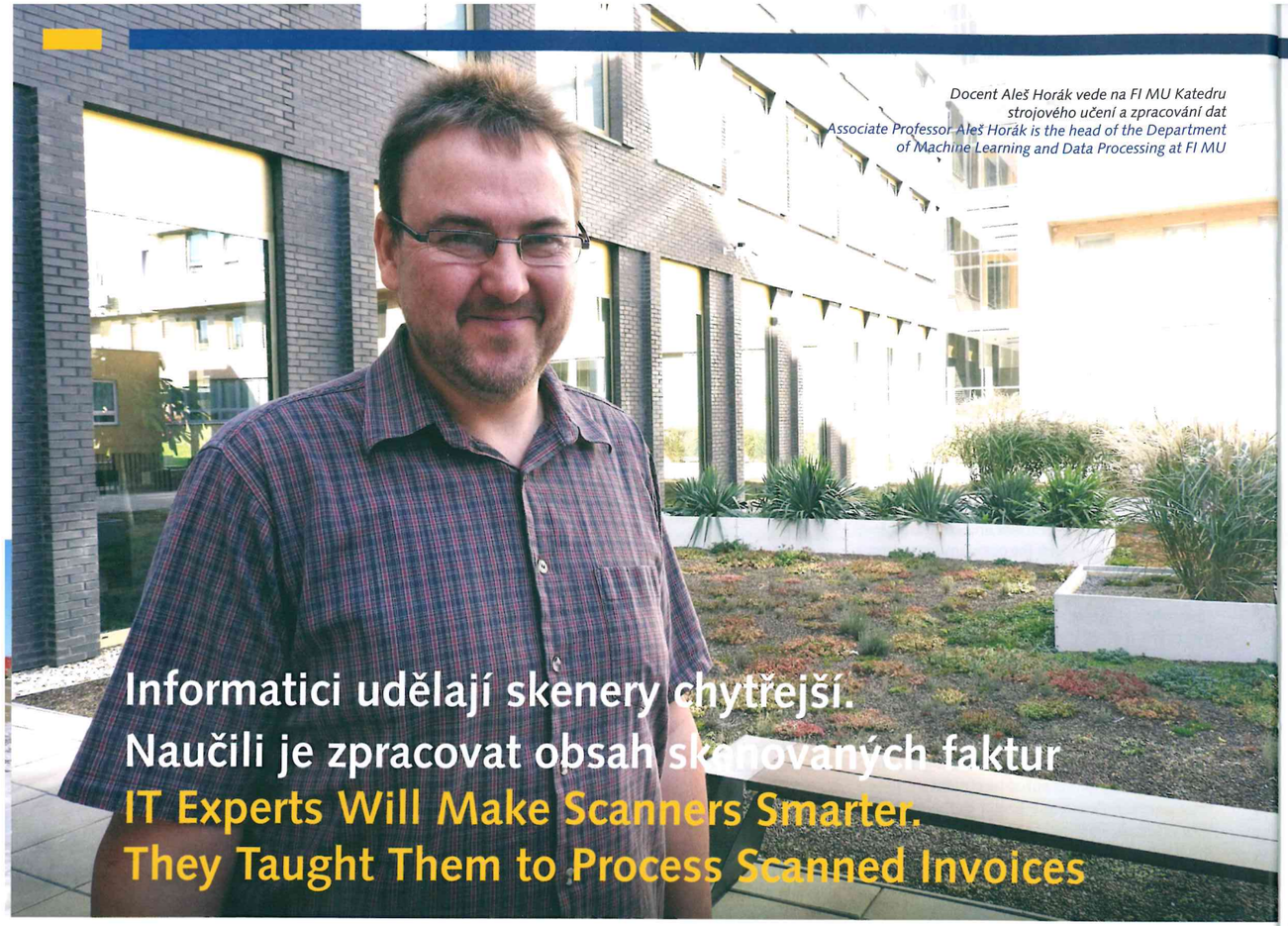


Docent Aleš Horák vede na FI MU Katedru strojového učení a zpracování dat
Associate Professor Aleš Horák is the head of the Department of Machine Learning and Data Processing at FI MU

Informatici udělají skenery chytřejší. Naučili je zpracovat obsah skenovaných faktur

IT Experts Will Make Scanners Smarter. They Taught Them to Process Scanned Invoices

Dokumenty naskenuje, označí a správně roztřídí podle zadaných kritérií. Na požádání pak poptávané informace a uložené doklady vyhledá – třeba tak, že přístroj jednoduše oslovíte, on vám odpoví a najde vše potřebné. Vítejte v kanceláři budoucnosti. Na extrakci informací ze skenovaných dokumentů, která by dnešní kopírky a tiskárny posunula k takzvaným „document management systems", pracuje i tým informatiků z Masarykovy univerzity. Na ověření jedné z technik využili podporu Proof of Concept.

It scans documents, marks them and sorts them according to given parameters. When asked, it searches the stored data and invoices – simply tell the machine to do so and it will find the required information. Welcome to the office of the future. A team of IT experts from Masaryk University is also tackling the task of extracting information from scanned documents which would push today's copying and printing machine closer towards so-called document management systems. To verify one of the methods they used the Proof of Concept support.

„Už delší dobu spolupracujeme se společností Konica Minolta, která se zabývá výrobou velkých skenovacích zařízení a tiskáren. Chtějí svá skenovací zařízení udělat inteligentnější, aby dokument nejen převedla na obrázek, ale aby i rozpoznala jeho obsah," popisuje docent Aleš Horák projekt OCR miner, tedy dolování dat ze skenovaných dokumentů.

Pod jeho vedením se tým odborníků z Fakulty informatiky zaměřil v Proof of Concept projektu na konkrétní úkol: extrakci údajů ze skenovaných faktur. Právě zvolený typ dokumentu byl pro vědecký tým výzvou. „Faktury jsou na extrakci konkrétních dat relativně složité. Když se totiž podíváme třeba na naše testovací data, což je tisícovka faktur od zhruba 50 dodavatelů z celého světa, vidíme, jak je formát faktur extrémně různorodý. Samozřejmě obsahují 10 až 20 základních typů informací, ale ty jsou uspořádané v naprosto nečekaných kombinacích, tvarech a formátech. Některé faktury jsou téměř umělecké dílo," hodnotí s nadsázkou Horák.

Díky finanční podpoře v rámci projektu zvládli výzkumníci během roku a půl vytvořit prototyp nástroje, který má za úkol rozpoznat konkrétní oblasti dat jako je třeba částka či adresa. „V rámci Proof of Concept vznikla sada na sebe navazujících nástrojů, každý z nich je specializovaný na jednu jazykovou technologii. V prototypu dosahujeme asi 80% úspěšnosti extrakce informací z faktury. A po přechodu na lepší OCR základ by se mohla úspěšnost systému ještě zvýšit," chválí si Horák.

Prototyp je zatím navržen pro češtinu a angličtinu. Firma Konica Minolta, která na projektu spolupracovala, by v budoucnu ráda pokryla více světových jazyků. „Zaměřují se samozřejmě primárně na jiné jazyky, než je čeština, ale firma má v Brně poměrně velké výzkumné centrum, z tohoto pohledu pro ně nástroj v češtině dává smysl. A nejen to, naše mateřština totiž patří, co se týče získávání informací ze struktury textu, ke složitějším jazykům. Zjevné je to například v porovnání s angličtinou, která je mnohem schematičtější a proto se v ní lépe vyhledává," vysvětluje Horák.

Úspěšnost řešení z dílny brněnských vědců je přitom srovnatelná s návrhy, které vznikly v zahraničí. „Z literatury známe pokusy o podobný úkol. Obvykle dosahují úspěšnosti mezi 70 a 90 % v případě angličtiny. My jsme použili trochu jiný přístup, zejména v zapojení technik analýzy jazykového obsahu. Během relativně krátkého času jsme dosáhli výsledků, které jsou podle literatury mezi těmi nejlepšími," dodává Horák.

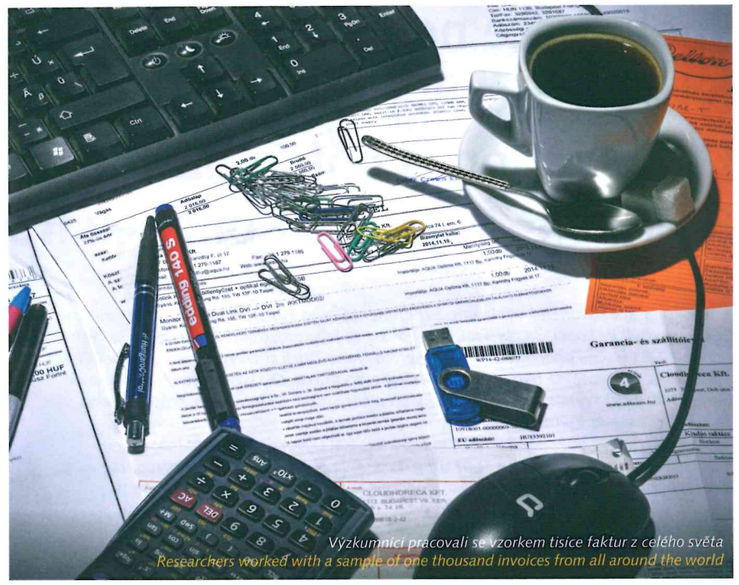
Bude na čem stavět

V projektu, který má udělat skenery a tiskárny „chytřejší", chtějí vědci pokračovat. Úkolem je rozšiřování portfolia dokumentů, které bude nástroj schopen rozpoznávat. „Nyní budeme pracovat na extrakci informací ze smluv. Předpokládáme, že kostra nástroje, která vznikla v rámci PoC, bude aplikovatelná i na jiné typy dokumentů, takže bychom na ní chtěli stavět," říká Horák.

Zavedení do praxe je úkolem firmy, Horák si proto netroufá říct, kdy přesně se tak stane, mohlo by jít řádově o jednotky let. Kromě vylepšení zařízení typově podobných těm, která jsou už dnes na trhu, směřuje celá oblast techniky mnohem dál. „Už dnes jsou kopírovací stroje vybavené počítačem a tím pádem jsou schopny pracovat s daty mnohem komplexněji, než to dělá klasická kopírka nebo skener. Vize je přetvořit kopírku na tzv. document management system, který dokumenty zpracovává komplexně včetně dat, kategorizace či inteligentního vyhledávání. Začínáme pracovat i nad tématem dotazovacích systémů pro komunikaci člověk-stroj. I ty mohou být někdy v budoucnu součástí těchto zařízení. Místo přes formulářové rozhraní by mohl uživatel s přístrojem komunikovat napřímo," popisuje Horák.

Člověk versus stroj

Během posledních pěti let se vývoj v oblasti strojového učení a zpracování dat rozběhl závratnou rychlostí. Kdo je dnes přesnější, člověk, nebo stroj? „V mnoha oblastech je stroj přesnější. Například ve zpracování velkých objemů dat, kterého prostě lidé nejsou schopni. Před časem jsme taky viděli počítač, který porazil jednoho z nejlepších lidských hráčů ve hře go. Někdy je ale otázka, co slovo „přesnější" vlastně znamená. Existuje třeba aplikace na rozpoznávání obsahu obrázků, která měla přesnost 99 %, lidská přesnost na testovací sadě byla 96 %. I když lidé v tomto případě za strojem na první pohled o tři procenta zaostávají, nelze to takto přímočaře interpretovat. Jak můžeme vědět, že má pravdu stroj, když se na obsahu těch několika procent obrázků lidé neshodnou?" uzavírá Horák. ■



Výzkumníci pracovali se vzorkem tisíce faktur z celého světa
Researchers worked with a sample of one thousand invoices from all around the world

"For quite some time now, we've been collaborating with Konica Minolta who develop and manufacture large scanning and printing machines. They want to make their scanners more intelligent so that they not only turn documents into digital images but also recognize their contents," says Assoc. Prof. Aleš Horák about the project OCR Miner, the aim of which is to mine data from scanned documents.

Under his direction, the team of experts from the Faculty of Informatics focused on a specific task within the Proof of Concept project: to extract data from scanned invoices. The type of document they chose posed quite a challenge. "As far as extracting specific data is concerned, invoices are relatively complicated. Let's have a look at our test data which consist of 1000 invoices from around 50 companies all over the world. One can see how extremely diverse the layouts of the invoices are. They still contain 10 to 20 basic types of information; however, these pieces of information are arranged in unexpected combinations, shapes, and formats. Some invoices are almost a work of art," says Mr. Horák jokingly.

Thanks to the financial support from the project, the researchers managed to come up with a prototype of a tool which recognizes the specific types of data such as addresses or sums of money. "Within the Proof of Concept, a set of tools was created each of which specializes in one language technology. The prototype extracts information from invoices with a success rate of around 80 %. After changing over to a better OCR base, the success rate could be even better," says Mr. Horák.

The prototype currently works with Czech and English. Konica Minolta who collaborated on the project would like to cover more world languages. "Naturally, their primary focus is on other languages than Czech; however, the company has a relatively large R&D centre in Brno and from this point of view, the tool makes good sense to them. Moreover, our mother tongue is one of the more complex languages which shows in extracting information from text as well. This is clear especially in comparison with English which is much more schematic and therefore easier to search in," explains Mr. Horák.

The success rate of the solution developed by the scientists from Brno is comparable with other tools developed abroad. "We're aware of several attempts at a similar task. Usually, they have a success rate between 70 and 90 % in the case of English. We took a slightly different approach, especially by integrating certain methods of language analysis. In a relatively short time we achieved, according to specialized literature, one of the best results there are," adds Mr. Horák.

Something to Build Upon

The scientists would like to continue working on this project that will make scanners and printers smarter. Another goal is to broaden the portfolio of documents the tool is able to recognize. "Now we're going to work on extracting information from contracts. We assume that the basic structure of the tool we have created within the PoC will be applicable to other types of documents as well," says Mr. Horák.

Implementing this technology in commercial products is the company's job. That's why Mr. Horák cannot say when it will happen but it could be in just a couple of years. It is not only about improving the devices similar to those that are already on the market – the entire technological field is developing rapidly. "Even nowadays copying machines are fitted with computers and therefore capable of processing data in a much more complex way than regular scanners or copying machines. The vision is to turn regular copying machines into so-called document management systems offering complex processing of documents including data analysis, categorization and intelligent searching. We've also started working on query systems for man-machine communication. One day, these may also be integrated into such devices. Instead of using forms interface, one could communicate with the machine directly," describes Mr. Horák.

Man Versus Machine

In the last five years, the development in the field of machine learning and data processing has sped up considerably. Who is currently more accurate, man or machine? "In many fields, it is the machine. For example in processing large volumes of data which people are simply not capable of. A while ago we saw a computer beat one of the best Go players in the world. However, sometimes it is not clear what "more accurate" means. For instance, there is an application for recognizing contents of pictures with the accuracy of 99 % whereas the human accuracy measured on a testing sample was only 96 %. Even though it might seem that people are 3 % less accurate than the machine, it is not that simple. How can we say that the machine is right if people can't agree on what it is that these 3 % of images actually show?" concludes Mr. Horák. ■

Text Iveta Zieglová,
foto Iveta Zieglová a Pixabay.com