Posudek *oponenta* diplomové práce

Jméno a příjmení autora posudku: **RNDr. Michal Kopecký, Ph.D.**

Jméno a příjmení autora práce: **Bc. Jakub Michalko**

Název práce: **Algoritmy detekce obchodních dokumentů podle šablon**

Vlastní text:

Cílem práce bylo navrhnout a implementovat program, který by dokázal klasifikovat elektronické dokumenty podle šablon, a na základě této klasifikace následně dokázal extrahovat data z dokumentu s ohledem na sémantiku, tedy na základě znalosti toho, že v dané oblasti se s největší pravděpodobností nachází adresa, telefon, a podobně.

Práce je obhajována podruhé, a od předchozí obhajoby autor zohlednil většinu připomínek. Text práce je nyní lépe formálně strukturován, a lépe odděluje formální popisy algoritmů od jejich konkrétní implementace. Dále přibylo uživatelské rozhraní, které umožňuje komfortněji dokumenty analyzovat, kontrolovat výsledek automatické anotace, a případně doplňovat ta data, která nebyla automaticky rozpoznána.

Jádrem řešení je volně šiřitelný OCR nástroj Tesseract. S jeho pomocí je dokument převeden z obrázku na text, který je následně anotován a doplněn o pozice výskytu slov v textu a tento meziprodukt je dále zpracován.

Oproti předchozí verzi přibyla i možnost přímějšího zpracování (rozumného, textovou vrstvu obsahujícího) PDF souboru, který je nyní možné zpracovat přímo bez nutnosti jej konvertovat na obrázek a ten zpět pomocí OCR na text.

Rovněž přibyla sekce s testy vytvořeného řešení v porovnání se SW projektem LANA, vytvořeným na MFF, a který využívá stejný OCR nástroj. Na základě testů je pak aplikace srovnána i s několika dalšími řešeními, které práce uvádí.

Hlavním přínosem oproti pouhému převodu na text je gramatikou řízené rozpoznávání typů údajů na stránkách na základě obsahu a umístěním bloku dat vůči jiným blokům s již určeným typem.

Celkově autor splnil zadání práce.

Doporučení k obhajobě:

Práci *doporučuji* k obhajobě.

Vynikající práce vhodná pro soutěž studentských prací	NE

Seznam soutěží studentských prací, viz http://www.mff.cuni.cz/studium/bcmgr/prace/

Pokud jste výše zaškrtli ANO, zdůvodněte prosím svůj návrh, případně uveďte konkrétní soutěž, pro kterou je práce vhodná (rámeček lze nechat prázdný, pokud za dostatečné zdůvodnění považujete text posudku):

V Praze dne: 23. 8. 2016

Podpis: