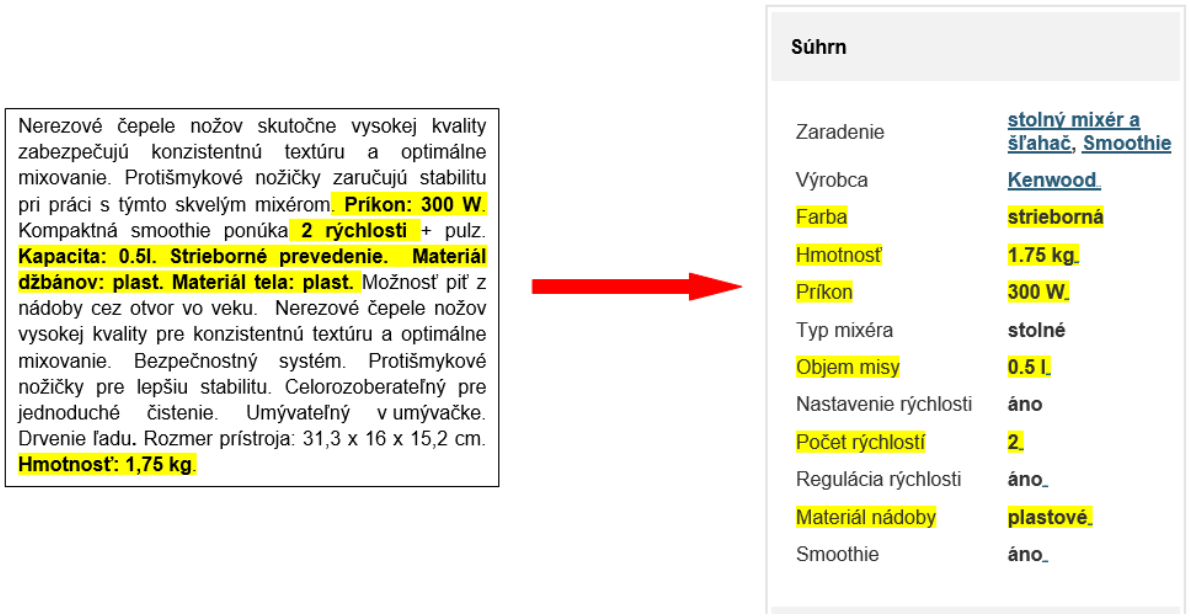
Získavanie atribútov objektov z ich popisu

Vedúci práce: RNDr. Peter Gurský, PhD.

Diplomová práca sa zaoberá extrahovaním informácii -atribútov objektov z voľného textu. Cieľom je navrhnutie takej sady metód, ktorá by automaticky extrahovala atribúty z popisu. Preferujeme návrhy metód, ktoré by vedeli extrahovať atribúty rôznych typov. Ako napríklad reťazový, číselný, atribút typu áno-nie. Pri návrhu metód môžeme využiť vopred extrahované štruktúrované dáta v podobe názvov, hodnôt, jednotiek atribútov.

Obr.1. znázorňuje príklad vstupu a výstupu metód na extrakciu. Vstupom je popis v neštruktúrovanej podobe produktu. Z tohto popisu chceme vybrať atribúty ako napríklad príkon, farba, materiál, hmotnosť a pod. Tieto atribúty následne zapisujeme do tabuľkovej formy



Obr1. ~~Ilustračný~~ príklad vstupu a výstupu extrakčných metód

Práca nadväzuje na bakalársku prácu s názvom Automatická extrakcia atribútov z popisu produktov, ktorej výsledkom boli návrhy metód pre rôzne typy atribútov. Navrhnuté boli metódy pre atribúty číselného, reťazového typu a typy áno-nie. Po testovaní týchto metód na vzorových popisoch produktov, nasledovala analýza metód . Porovnávali sa požadované hodnoty oproti hodnotám, ktoré sa získali metódami extrakcie. Analýzou sa zisťovali taktiež dôvody, ktoré ovplyvňovali neúspešnosť metód po stránke úplnosti a presnosti. Medzi nich patrilo:

1. Nájdenie atribútu, ktorý sa v popise reálne nenachádza
2. Výskyt chyby v popise
3. Neprítomnosť názvu atribútu v slovníku
4. Zmena tvaroslovia hodnoty alebo názvu atribútu

5. Vynechanie slova alebo slov z viacslovných pomenovaní atribútov
6. Rozdelenie slov viacslovného pomenovania
7. Použitie skratky slova názvu

Cieľom diplomovej práce je navrhnúť metódy, ktoré by chyby minimalizovali. V návrhoch ~~by~~ chceme použiť metódy pre spracovanie prirodzeného jazyka, ktoré sa venujú lemmatizáciu, stemming a pod. Uvažujeme aj nad využitím pripraveného tvaroslovníka a tiež skrytých Markových modelov.

Navrhnuté metódy budeme implementovať použitím programovacieho jazyka Java. Ďalším krokom v práci bude vytvoriť dostatočne veľkú dátovú sadu popisov a atribútov, ktoré sa nachádzajú v daných popisoch, aby sme mohli metódy testovať a analyzovať ich korektnosť, úplnosť a zistiť prípadné zlyhania metód.

Dátovú sadu získame vytvorením webového rozhrania, v ktorom sa budú manuálne extrahovať atribúty z pripravených neštruktúrovaných textov objektov a následne sa tieto atribúty budú uchovávať v databáze na testovanie metód.

Do literatúry k diplomovej práci patria články, ktoré sa zaoberajú named-entity recognition (NER), pretože táto oblasť je blízka nášmu výskumu. NER je extrakcia informácií na základe identifikácií a klasifikovaní zmienok v texte o ľuďoch, organizáciách, miestach a iných entitách. Zvyčajným prístupom metód NER je označenie slov v testovacou texte na základe anotovaných entít vo veľkom tréningovom súbore textov. Na zvládnutie NER bolo navrhnutých viacero učiacich algoritmov ako je Hidden Markov Model, rozhodovacie stromy, Support Vector Machines, Conditional Random Fields. Touto oblasťou sa venujú články *Learning multilingual named entity recognition from Wikipedia, A survey of named entity recognition and classification, High-Performance Learning Name-finder,Named Entity Recognition Using Support Vector Machine for Filipino Text Documents*

Ďalšou podobnou oblasťou je extrakcia názvov/entít. Cieľom tejto extrakcie je automatická extrakcia relevantných názvov z textu založených na slovách zo slovníka domenovo podobných názvov. Príkladom môže byť metóda C-value/NC-value, ktorá je opísaná v článku *The C-value/NC-value Method of Automatic Recognition of Multi-word Terms.*

Keďže uvažujeme o využití tvaroslovníka v našej práci, v literatúre je zaradená aj diplomová práca *Syntaktická analýza slovenskej vety pomocou Tvaroslovníka*, ktorá sa zaoberala daným slovníkom a jeho využitím v analýze vety.

Odporúčaná literatúra

1. J. NothMan at al.: Learning multilingual named entity recognition from Wikipedia. Artificial Intelligence 194 (2013) 151–175
2. D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvisticae Investigationes 30 (2007) 3–26
3. D. M. Bikel et al.: Nymble: a High-Performance Learning Name-finder. In ANLP-97, Washington, D.C., pp. 194 – 201, 1997.

4. J. M. Castillo et al.: Named Entity Recognition Using Support Vector Machine for Filipino Text Documents. International Journal of Future Computer and Communication, Vol. 2, No. 5, October 2013
5. K. Frantzi, S. Ananiadou, J. Tsujii: The C-value/NC-value Method of Automatic Recognition of Multi-word Terms. In proceedings of ECDL, pp. 585-604. ISBN 3-540-65101-2, 1998
6. Jana Hiľovská: Syntaktická analýza slovenskej vety pomocou Tvaroslovníka. Diplomová práca PF UPJŠ 2017