INFORMATION RETRIEVAL FROM SPOKEN DOCUMENTS

Michal FAPŠO, Master Degree Programme (1)
Dept. of Computer Graphics and Multimedia, FIT, BUT
E-mail: xfapso00@stud.fit.vutbr.cz

Supervised by: Ing. Petr Schwarz

ABSTRACT

This paper describes a designed and implemented system for efficient storage, indexing and search in collections of spoken documents that takes advantage of automatic speech recognition. As the quality of current speech recognizers is not sufficient for a great deal of applications, it is necessary to index the ambiguous output of the recognition, i. e. the acyclic graphs of word hypotheses — recognition lattices. Then, it is not possible to directly apply the standard methods known from text-based systems. The paper discusses an optimized indexing system for efficient search in the complex and large data structure that has been developed by our group.

1 ÚVOD

Rozpoznávač reči dokáže previesť zvukový záznam na text, v ktorom by sme mohli priamo vyhľadávať štandardnými metódami. Problém tohoto prístupu ale spočíva v tom, že rozpoznávač reči nemá nikdy stopercentnú úspešnosť. Niektoré slová vyskytujúce sa v zázname budú teda nesprávne rozpoznané, a preto ich textový vyhľadávač nedokáže nájsť.

Systém popísaný v tomto príspevku však pracuje na odlišnom princípe. Nevyhľadáva v textovom výstupe rozpoznávača, ale v tzv. grafoch hypotéz. V nich sa nachádza okrem prepisu označeného rozpoznávačom za najpravdepodobnejší i množstvo iných menej pravdepodobných hypotéz. Takýmto spôsobom teda nájdeme i slová, ktoré vo vyššie popísanom jednoduchšom systéme vyhľadať nešlo.

Vzhľadom na veľký objem dát a ich zložitú štruktúru (orientované grafy hypotéz) je nutné vytvoriť štruktúru indexov, ktoré umožnia rýchle vyhľadávanie a prístup k dátam [1].

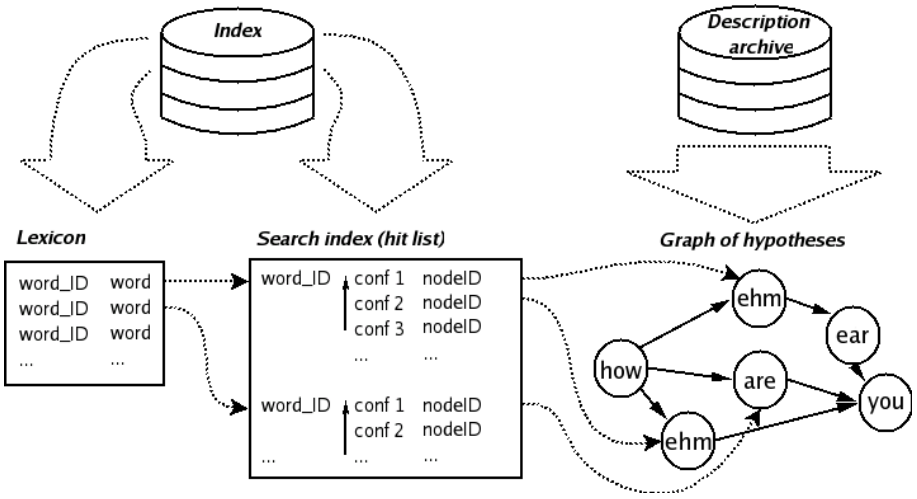
2 ROZPOZNÁVAČ REČI

Pre účely vyhľadávania môžeme použiť dva typy rozpoznávačov: slovný alebo fonémový. Fonémový rozpoznáva základné rečové jednotky - fonémy (napr. SEARCHING =>

s er ch ih ng), zatiaľčo slovný rozpoznáva celé slová, pričom je obmedzený slovníkom (cca. 50 000 slov). Aj keď je slovný rozpoznávač vďaka použitiu slovníka a jazykového modelu [2] pri rozpoznávaní úspešnejší, nedokáže rozpoznať slová, ktoré v slovníku nemá (vlastné mená, odborné výrazy, ...). Ak nechceme byť pri vyhľadávaní obmedzení slovníkom, musíme vyhľadávať vo výsledku fonémového rozpoznávača. Kombináciou oboch typov rozpoznávačov však docielime vyššiu presnosť i robustnosť. Slová nachádzajúce sa v slovníku vyhľadáme vo výsledku slovného rozpoznávača a ostatné slová prevedieme na ich fonémový prepis [3] a vyhľadáme vo výsledku fonémového rozpoznávača.

3 ŠTRUKTÚRA SYSTÉMU

Po rozpoznaní reči a vygenerovaní slovných a fonémových grafov hypotéz oboma rozpoznávačmi nasleduje prevedenie týchto grafov do binárnej podoby a indexovanie (Obr. 1).



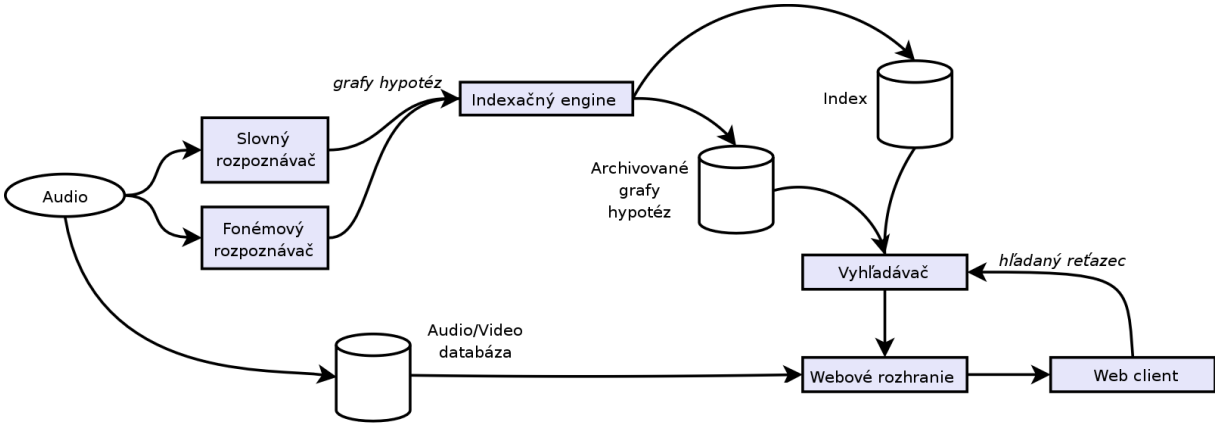
Obrázek 1: Zjednodušená schéma štruktúry indexov

V tejto fáze môžeme začať vyhľadávať. Samotný vyhľadávač pracuje ako server, ktorému klienti posielajú požiadavky a ten im posiela výsledky hľadania.

V rámci európskeho projektu AMI [4] bola vyvinutá aplikácia na interaktívne prehliadanie mítingov (JFerret), ktorá bola rozšírená o možnosť vyhľadávania pomocou komunikácie s vyhľadávacím serverom.

4 PRIEBEH VYHĽADÁVANIA

Pre každé slovo z hľadanj frázy sa pomocou slovníka zistí identifikátor hľadaného slova, pomocou ktorého sa z reverzného indexu [1] načítajú informácie o všetkých výskytoch daného slova v indexovaných dátach. Následne sa vyberú skupiny výskytov jednotlivých slov v čase tak, aby vyhovovali zadanej fráze. Rozdiel oproti vyhľadávaniu v textových dátach je v tom, že v grafe hypotéz má každá hypotéza (slovo, foném) priradenú pravdepodobnosť akou ju ohodnotil rozpoznávač. Nájdené skupiny hypotéz zoradíme podľa tejto pravdepodobnosti. Nevieme však, či pre každú skupinu existuje v grafe



Obrázek 2: Schéma implementovaného systému pre vyhľadávanie

cesta cez jej hypotézy. Preto musíme pre každú skupinu prejsť príslušnú časť grafu a zistiť, či cesta medzi hypotézami existuje. Podobným spôsobom zistíme i najpravdepodobnejší kontext nájdeného slova alebo frázy.

5 ZÁVER

Popísaný systém je implementovaný a čoskoro sa plánuje jeho experimentálne nasadenie na vyhľadávanie v záznamoch z prednášok.

POĎAKOVANIE

Tento príspevok vznikol za podpory EC projektu Posilnenej skupinovej interakcie (AMI) č. 506811 a grantu GAČR 102/05/0278.

REFERENCE

[1] Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Science Department, Stanford University

[2] Igor Szöke et al.: Comparison of Keyword Spotting Approaches for Informal Continuous Speech, In Proc. Eurospeech 2005, Lisabon, Portugal, September 2005.

[3] Bosch, A., Daelemans, W.: Data-Oriented Methods for Grapheme-to-Phoneme Conversion, Tilburg university, 1993

[4] Augmented Multi-party Interaction, www.amiproject.org

[5] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book, Cambridge University Engineering Department, 2002