Studentská konference Fakulty mechatroniky, informatiky a mezioborových studií
7. červen 2012, Liberec, Česká republika

Automatická sumarizace textových dokumentů

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky a mezioborových studií

Bc. Michal Rott
Ing. Petr Červa Ph.D.
Ústav Informačních technologií a elektroniky

Abstrakt

Automatic summarization is a linguistic discipline, which is used to create a summaries from large texts.

Heuristic summarization method, Luhn summarizer and Method of Latent semantic analysis have been implemented.

All implemented methods were evaluated with ROUGE evaluation tool.

Online available summarizers, which were able to process Czech language, has been evaluated.

Simple website was developed to be able to summarize Czech articles.

This website is now available on adress: http://nashida.ite.tul.cz.

Cíle

Nastudovat principy sumarizace textu.
Implementovat vybrané metody sumarizace.
Vytvořit sumarizační systém pro textové dokumenty.
Provést evaluaci implementovaných metod pro český a anglický jazyk.
Porovnat implementované metody s běžně dostupnými sumarizátory.

Problematika

Počátky - 60. léta 20. století
- nedostatečná kapacita pro uložení digitalizovaných textů
- náhrada textů souhrny
- chybějící souhrny doplněny pomocí automatické sumarizace

Dnes - „nekonečná“ úložná kapacita
- denně vzniká ohromné množství textů
- přehlcení informacemi
- nutnost rozhodnout se co je pro nás důležité číst

Rozsah souhrnu - vyjádřen kompresním poměrem

Informativní souhrn:
- obeznámení se s tématikou textu
- obsahuje 20 až 30 % původního textu

Indikativní souhrn:
- určuje zda má smysl číst původní text
- obsahuje do 10 % původního textu

Extrakt
- nejčastější produkt automatické sumarizace
- neupravené věty extrahované z původního textu

Metody sumarizace

Heuristická metoda
- využití četnosti termů textu (term = část textu označující věc, děj, ...)
- odstranění termů ze Stoplistu (Stoplist = nejběžnější termy jazyka, nenesou žádný význam pro obsah věty)
- skóre věty vypočteno jako součet četností termů
- do souhrnu jsou vybrány věty s nejvyšším skóre.

Luhnův sumarizátor
- potřeba natrénovaného slovníku inverzní dokumentové frekvence [1]
- IDF skóre využito k vážení termů (není nutný Stoplist)

$$skóre(t,d) = tf(t,d) * idf(t,D) = tf(t,d) * \log \frac{|D|}{|d \in D : t \in d|} \quad (1)$$

- IDF skóre natrénováno na korpusu dokumentů
- na kvalitě korpusu závisí i kvalita generovaného souhrnu

Latentní sémantická analýza
- řešení sumarizace pomocí algebraické metody
- věty a termy textu jsou mapovány do matice A [2]
- sloupce - věty textu; řádky - termy; hodnota - frekvenci termu ve větě
- matice A rozložena pomocí singulární dekompozice

$$A = U * \Sigma * V^T \quad (2)$$

- vybrány věty s největší normou sloupcových vektorů v matici V^T

Evaluace metod

Evaluace byla provedena pomocí programu **ROUGE** [3].
- využíván na konferencích TAC - Text Analysis Conference
- umožňuje efektivně evaluovat extraktivní metody sumarizace
- požaduje pro evaluaci referenční souhrny vytvořené anotátory
- vytvořena databáze referenčních souhrnů od *11 anotátorů*
- vybráno *25 článků* ze serverů aktualne.cz a novinky.cz (témata článků: kultura, ekonomie, domácí a zahraniční události)
- vypočteny čtyři druhy skóre určující míru podobnosti referenčních a automaticky generovaných souhrnů
- vybraná skóre: *nejdelší společná subsekvence (LCS)* počet shodných *unigramů*, *bigramů, skip-bigramů*
- skóre vyjádřeno pomocí hodnot úplnosti a přesnosti

Úplnost reprezentuje kolik procent sekvencí vět vybraných anotátory vybral i sumarizátor.

Přesnost reprezentuje kolik procent sekvencí vět vybraných sumarizátorem vybrali i anotátoři.

Vyhodnocení experimentů

Evaluovány byly všechny implementované metody ve variantách s klíčovými slovy textu a bez nich. Vyzkoušen byl i vliv ohebnosti jazyka na výsledek sumarizace. Také byly evaluovány volně dostupné sumarizátory, které dokázaly zpracovat český jazyk.

	unigramy		bigramy		LCS		skip-bigramy	
metoda	úplnost	přesnost	úplnost	pčesnost	úplnost	přesnost	úplnost	přesnost
Heuristická	65,0	60,8	51,3	48,0	23,5	43,2	36,8	34,5
Luhnova	72,2	60,8	58,7	49,7	25,9	42,8	44,1	32,0
LSA	75,2	60,9	62,4	50,5	27,4	43,5	47,4	31,8
OTS	56,6	62,6	42,1	47,8	19,9	44,2	30,8	40,4
T4N	73,3	57,2	59,9	46,7	23,9	41,3	46,2	28,9
LSA - EN	69,4	60,4	54,9	47,6	24,6	46,9	40,1	30,4

Poznatky:
- Pro český text není potřeba využívat náročnou LSA metodu. Luhnův sumarizátor produkuje stejně dobré souhrny.
- Po lemmatizaci českého textu jsou výsledky lepší než pro neohebný anglický jazyk → lemmatizace je důležitá při procesu sumarizace.
- Běžně dostupné sumarizátory nedosáhly tak dobrých výsledků jako metoda Luhnova nebo LSA. Výjimkou je sumarizátor T4N (placený).

Referece

[1] Luhn. H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 1958.

[2] Josef Steinberger and Karel Ježek. Text summarization and singular value decomposition. *Advances in Information Systems*, 2005.

[3] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *In Proceedings ACL workshop on Text Summarization Branches Out*, 2004.

Autor

Bc. Michal Rott
e-mail: michal.rott@tul.cz
www: http://nashida.ite.tul.cz

Prezentace této práce byla podpořena z projektu SGS 2012.