Automatická sumarizace textových dokumentů

Michal Rott, Petr Červa

Abstrakt

Tento projekt se snaží lidem usnadnit práci s informacemi vytvářením souhrnů textů, které tyto informace obsahují. V rámci výzkumu byly zkoumány metody vytvářející z rozsáhlých článků extrakty, které obsahují zásadní informace těchto článků. Byly nastudovány sumarizační metody heuristické a statistické využívané v počátcích digitalizace textů, ale i moderní metody analyzující texty hlouběji. Implementované metody byly také vyhodnoceny na základě referenčních souhrnů.

Úvod

Dnes vznikají denně tisíce článků a dokumentů. Přečíst všechny tyto texty ovšem není v lidských možnostech a proto začaly být nasazovány sumarizační systémy, které z těchto článků vytváří souhrny.

První sumarizační metody se objevily v 60. letech, kdy se začaly v knihovnách digitalizovat dokumenty. Ovšem tehdejší kapacita datových úložišť nebyla dostačující pro uložení všech dokumentů, takže místo celých dokumentů byly elektronicky uloženy jen jejich souhrny.

Hlavním cílem práce bylo vytvoření sumarizačního systému, který by umožňoval vytváření extraktů dokumentů na základě různých sumarizačních metod a vyhodnotit podobnost souhrnů vytvořených tímto systémem se souhrny vytvořenými anotátory.

Experimenty a metody

Pro implementaci byly vybrány celkem tři metody: Heuristická metoda, statistická metoda fungující na principu ohodnocení vět dokumentů na základě frekvence jejich termů a inverzní dokumentové frekvence těchto termů a poslední byla implementovaná metoda využívající Latentní sémantickou analýzu textu. Termy jsou rozuměny části textu, které popisují jeden reálný jev (osobu, věc, děj,...).

Heuristická metoda využívá povrchních znalostí o textu k jeho sumarizaci. Těmito znalosti je míněna například četnost výskytu termu, poziční význam termů, termy zvýrazňující význam věty, nebo termy bez významu pro souhrnu (StopList) [1].

Luhnův sumarizátor zavádí do problematiky hodnocení termů znalosti o jejich výskytech v jazyku. Tato metoda násobí frekvenci termu jejich inverzní dokumentovou frekvencí: Score(t,d)=tf(t,d)*idf(t,D)

Latentní sémantická analýza je metoda založená na myšlence, která je využívána Latentním sémantickým indexováním [2]. Na začátku procesu sumarizace je vytvořena matice příznaků vět. Tato matice je pomocí singulární dekompozice rozložena na matice: $A = U\Sigma V^T$. Pomocí matic Σ a V^T jsou ohodnoceny věty dokumentu a vybrány věty jejichž norma vektoru je největší [3].

Pro **vyhodnocení kvality** implementovaných metod byly navrženy tři experimenty. Vyhodnotit kvalitu na základě referenčních souhrnů. Porovnat sumarizátor se sumarizátory dostupnými na Internetu. Porovnat sumarizaci ohebného a neohebného jazyka.

Za účelem vyhodnocení sumarizačních metod byl použit evaluační software ROUGE [4], který pro evaluaci systémů využívá základní metriky měření ko-sekvencí vět: nejdelší společná subsekvence, počet společných n-gramů a počet společných skip-n-gramů.

Pro realizaci experimentů byla vytvořena databáze českých článků, které byly předloženy anotátorům. Anotátoři měli za úkol označit 25 % vět v článku čísly podle toho, jak si myslí, že jsou věty důležité.

Z těchto článků byly vytvořeny také automaticky generované souhrny a pomocí nich byly implementované metody hodnoceny.

Výsledky a diskuze

Výsledky evaluace jsou zobrazeny v procentech. Úplnost (recall) značí kolik procent vybraných vět mělo být skutečně vybráno, přesnost (precision) kolik procent vět vybraných anotátory vybral i sumarizátor. Byly vyzk̃oušeny i již existující systémy a dva z nich jsou uvedeny zde: konzolový Open Text Summarizer a webový Tools4Noobs.

Tabulka 1: Výsledky evaluace metod v procentech

	unigramy		bigramy		LCS		skip-bigramy	
	R	P	R	P	R	P	R	P
Heur.	65,0	60,8	51,3	48,0	23,5	43,2	36,8	34,5
Luhn	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	26,9	41,3	46,2	28,9
LsaEN	69,4	60,4	54,9	47,6	24,6	46,9	40,1	30,4

Závěr

Během práce byly implementovány tři sumarizační metody a tyto metody byly pomocí softwaru ROUGE evaluovány. Sumarizační metody umožňují vytvářet souhrny hlavně českých textů, ovšem některé lze velmi jednoduše rozšířit o možnost sumarizovat texty v jiném jazyce. Byly porovnány výsledky implementovaných metod s výsledky běžně dostupných sumarizátorů a implementované metody dopadly velmi dobře. Dále byly porovnány souhrny lemmantizovaných českých článků a anglických překladů původních článků. Výsledky všech experimentů jsou uvedeny v tabulce 1.

Poděkování

Chtěl bych poděkovat vedoucímu diplomové práce za spolupráci a také bych chtěl poděkoval všem anotátorů, kteří mi pomohli vytvořit referenční souhrny, bez kterých bych nemohl provádět evaluaci.

Reference

[1] Luhn. H. P. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, April 1958.

[2] Gong Yihong and Liu Xin. Generic text summarization using relevance measure and latent semantic analysis. *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.

[3] Josef Steinberger and Karel Ježek. Text summarization and singular value decomposition. *Advances in Information Systems*, Springer Berlin / Heidelberg, 2005.

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