Multi-dokumentová sumarizace novinových článků

Jaromír Novotný[1]

Abstrakt

Práce se zabývá multi-dokumentovou sumarizací novinových článků. Slovo sumarizace bude v této práci používáno jako pojem procesu vytváření zkrácené verze textu, která informativně odpovídá originálnímu textu. Toto jasné stanovení významu je dáno kvůli dalším možnostem použití. Výsledná zkrácená verze textu se též může nazývat sumarizace, a proto bude zde používán pojem souhrn. Aby bylo možné vysvětlit multi-dokumentovou sumarizaci novinových článků je v prvé řadě důležité vysvětlení pojmů sumarizace a souhrn. Sumarizace je proces získávání nejdůležitější informace ze zdroje (nebo zdrojů), pro vytvoření zkrácené verze pro konkrétního uživatele (nebo více uživatelů) a pro konkrétní úlohu (nebo úlohy). Sumarizační metody lze dělit různými způsoby, záleží například na tom, jak chceme danou úlohu řešit nebo na vstupních datech. V základu lze dělit metody na manuální nebo automatické. Určitý tvar vstupu je v mnoha případech pro automatické metody klíčový.

Sumarizace novinových (zpravodajských) článků vzniká jako důsledek velkého přírůstku zpravodajských článků na internetu. Jelikož v takovém množství je pro čtenáře velmi obtížné efektivně shromažďovat pro něj důležité informace, nastupuje zde automatická sumarizace, která nahrazuje celé články kratšími verzemi se stejným přínosem informací. Tudíž čtenáři nemusejí číst celé články, ale pouze nejvíce informativní části z nich (souhrny).

Jak může být z názvu patrné, vstupem vybrané multi-dokumentové automatické metody není pouze jeden vstupní text (dokument), ale několik textů (dokumentů) najednou. Vstup tedy může být reprezentován jedním nebo ve většině případů i více soubory obsahující texty (články). Většina automatických sumarizačních metod pro zpracovávání multi-dokumentů předpokládá, že články (dokumenty) budou odpovídat jednomu tématu nebo určitému časovému období (tedy budou zaměřeny na určitou událost). V případě, že tomu tak není, je nutné pro většinu automatických sumarizačních metod vstup roztřídit a to např. automatickou metodou na bázi shlukování.

Pro možnost otestování alespoň některých automatických sumarizačních metod na dané novinové články, byly vybrány tři metody (implementace těchto metod byla provedena v programovacím jazyce Python). Metoda založená na latentní sémantické analýze je první zástupce. Druhý zástupce je metoda založená na váze středů shluků. A jako poslední je metoda založená na nezáporné maticové faktorizaci a K-means shlukování.

[1] student navazujícího studijního programu Aplikované vědy a informatika, obor Kybernetika a řídící technika, specializace Umělá inteligence, e-mail: fallout7@students.zcu.cz

Aby bylo možné porovnat použitelnost těchto metod v praxi, je potřeba ohodnotit výsledné souhrny. Ohodnocování výsledků sumarizačních metod slouží k posouzení rozdílů kvality mezi jednotlivými sumarizačními metodami. Samozřejmě můžeme mluvit o ohodnocování, které provádí člověk na základě svých vlastních zkušeností. Výsledky takového ohodnocení jsou u každého jedince jiné, a kvůli tomu nejsou dostatečně přesné ke stanovení závěrů o nejlepší metodě. Zatímco automatické ohodnocování výsledků sumarizačních metod je prováděno danou metodou řízenou algoritmem. Tedy při správném zvolení metody již může být stanoven závěr o kvalitě jednotlivých automatických sumarizačních metod.

Všechny metody, které byly ozkoušeny lze použít jak na vstupy (články) z daného časového období tak na vstupy se stejným tématem. Díky tomuto ohodnocení byl učiněn závěr o kvalitě jednotlivých sumarizačních metod. Příklad jedné tabulky se shrnutými výsledky je uveden níže (viz tab. 1). Z výsledků lze říci, že nejlepší metoda pro použití v praxi je metoda založená na NMF a K-means shlukování.

	Metody			
	LSA+LexRank	LSA-délka vět	MEAD	NMF+K-means
téma 1	0.1263	0.1483	0.0974	**0.1492**
téma 2	0.2070	**0.2164**	0.0768	0.1340
téma 3	0.1409	0.1603	0.0794	**0.2044**
téma 4	0.1475	0.1523	0.0527	**0.1672**
téma 5	0.1348	0.1170	0.1519	**0.1648**
téma 6	0.0540	0.0947	0.0841	**0.1725**
téma 7	0.0834	0.0819	0.0464	**0.0970**
téma 8	0.0970	0.1622	0.0991	**0.2635**
téma 9	0.0500	0.1570	0.0639	**0.1823**
téma 10	0.0872	0.1265	0.0838	**0.1873**
téma 11	0.0618	**0.2031**	0.0965	0.2014
téma 12	0.1227	0.1694	0.1715	**0.2425**
téma 13	0.1060	0.1543	0.0314	**0.1785**
téma 14	0.0936	0.1280	0.1646	**0.1898**
téma 15	0.1473	0.2328	0.0479	**0.2642**
Celkový průměr	0.1106	0.1536	0.0898	**0.1866**

Tabulka 1: Průměry hodnot f-skóre (k určitým tématům) v případě použití ROUGE-2 ohodnocení

Literatura

Dragomir R. Radev, Hongyan Jing, Małgorzata Stys. *Centroid-based summarization of multiple documents***.** Information Processing & Management. 2004, číslo 40, Issue 6, strany 919-938, ISSN 0306-4573.

Sun Park, Ju-Hong Lee, Deok-Hwan Kim and Chan-Min Ahn. *Multi-document Summarization Based on Cluster Using Non-negative Matrix Factorization.* Jan van Leeuwen et al. (Eds.): SOFSEM 2007.

Josef Steinberger, Karel Ježek. *Evaluation Measures For Text Summarization***.** 28 vyd. Computing and Informatics. 2009. Plzeň: University of West Bohemia in Pilsen.

Chin-Yew Lin. *Looking for a Few Good Metrics: Automatic Summarization Evaluation — How Many Samples Are Enough?*. Proceeding of NTCIR-4, April 2003– June 2004.

Karel Ježek , Josef Steinberger. *Sumarizace textů*. Mikulov: DATAKON 2010, 16-19. 10. 2010.

Günes Erkan , Dragomir R. Radev. *LexRank: Graph-based Lexical Centrality as Salience in Text Summarization*. Ann Arbor: University of Michigan, 2004.

I. Mani, D. House, G. Klein, L. Hirschman, T. Firmin, B. Sundheim. *The TIPSTER SUMMAC Text Summarization Evaluation*