Automatická expandizace textů

Josef Steinberger & kolektiv Text-mining group
Západočeská univerzita v Plzni

Abstrakt: Automatická expandizace textů je nová průlomová teorie o generické tvorbě textů. Přináší nesčetné výhody všem spisovatelům, úředníkům a také vědeckým pracovníkům. Náš tým prozkoumal výhody, které by expandizační systém přinesl právě vědeckým pracovníkům, kteří jsou zahlceni prací a nestačí publikovat požadovaný počet článků. Výhody takového systému nás natolik zaujaly, že jsme se ihned pustili do vývoje expandizačního systému. Výsledkem výzkumu jsou dvě nové metody – náhodný a řízený výbuch slova. Dále prezentujeme dosažené výsledky testování vyvinutého systému a budoucí směry vývoje.

1. Úvod

Automatická sumarizace textů je vědecká oblast, jejíž důležitost se zvyšuje s rostoucí velikostí internetu a digitálních knihoven. Pokud uživatel listuje dokumenty v digitální knihovně, je nutné mu předložit jejich krátké výtahy (abstrakty nebo extrakty), které mu pomohou vybrat ten nejvhodnější dokument.

My jsme se však zaměřili na zcela opačnou a zatím neprobádanou oblast vědy, kterou je expandizace, tedy rozšiřování, textů. S požadavkem na expandizační systém se stále setkáváme ve vědecké praxi. Vědci (a tedy i doktorandi) jsou zavaleni prací a ještě k tomu musí každý rok publikovat několik článků, což je velice náročné. Expandizační systém, který by umožňoval takové články generovat, by přišel jistě vhod. Rozhodli jsme se tedy daný problém prozkoumat.

Cílem je vygenerovat článek o předem daném počtu stran ze znalosti pouhého abstraktu, ba pouze z klíčových slov. Výsledkem našeho bádání jsou dvě metody – náhodný výbuch slova a řízený výbuch slova. Tyto metody jsou obě založeny na korpusu textů. Popíšeme je v následujících kapitolách. Dále zde uvádíme dosavadní výsledky testování na různých kolekcích a směry dalšího výzkumu.

2. Náhodný výbuch slova

Jedná se o velice jednoduchou metodu, která je schopna vygenerovat větu z jediného slova. Základem je vytvoření korpusu textů z určité domény (například *Umělá inteligence*). Vstupem této metody mohou být buď klíčová slova nebo abstrakt, ze kterého získáme seznam (**α**) významových slov (vyloučením stopslov a převedením slov do základního tvaru). Slovům v seznamu nejprve přiřadíme podle četnosti počáteční ohodnocení. Seznam potom seřadíme (dle ohodnocení). Následuje řetězovitý proces, která probíhá podle následujícího algoritmu:

1. vybere se slovo z čela seznamu **α**

2. provede se náhodný výbuch slova vybraného v kroku 1.
   - z korpusu textů se vybere náhodná věta, která slovo obsahuje[1]
   - tato věta se vloží do generovaného textu
   - slovo vybrané v kroku 1. se odstraní ze seznamu **α**

3. z věty vybrané v kroku 2 získáme seznam slov **β** (opět vyloučíme stopslova a zbylá slova převedeme do základního tvaru)

4. spojí se seznamy **α** a **β**[2]**,** výsledek označíme opět **α** a seřadíme jej podle četnosti

5. pokud je článek již dostatečně dlouhý, proces končí, pokud ne, skáče se na krok 1.

Tento proces je velice jednoduchý, ale přesto poskytuje velice dobré výsledky (viz testování). Avšak setkáváme se zde také s několika problémy (například „Pořadový efekt" – viz závěr).

3. Řízený výbuch slova

Tato metoda je dokonalejší verzí náhodného výbuchu slova Předchozí metoda vykazuje solidní výsledky, ale je schopna vygenerovat pouze nestrukturovaný text. U vědeckých článků však potřebujeme určité rozčlenění – například:

1. Úvod
2. Hlavní téma
3. Výsledky
4. Závěr

Toto umožňuje řízený výbuch slova. Základ algoritmu je totožný s náhodným výbuchem slova, avšak udržujeme pro každou část (např. Úvod) vlastní seznam. Všechny seznamy se inicializují stejným vstupem (klíčová slova nebo významová slova abstraktu). Dále se zde mění krok 2. – z korpusu textů se opět vybere náhodná věta, která slovo z čela seznamu obsahuje, ale věta se musí vyskytovat v části, kterou právě generujeme. (Tedy pokud právě generujeme závěr, vybírá se pouze z vět, vyskytujících se v závěrech článků korpusu.) Generování výsledků také nepředstavuje velký problém – k nalezeným hodnotám se připočítá náhodný koeficient. Ale toto nepovažujeme za převratnou novotu.

4. Výsledky testování

Vytvořený expandizační systém jsme testovali na následujících kolekcích textů:

- kolekce katedrálních publikací – vědecké články (převážně v angličtině)
- kolekce erotických textů vytvořená Ing. Tesařem (výzkum protizákonné tématiky) – anglicky

Hlavním cílem výzkumu bylo generování vědeckých článků. Proto jsme nejprve sestavili kolekci katedrálních publikací (celkem 824 textů). Pro každý článek jsme vygenerovali:

- 10 článků metodou náhodného výbuchu, kde vstupem byla klíčová slova (metoda ﻝ)
- 10 článků metodou náhodného výbuchu, kde vstupem byl abstrakt (metoda ﻙ)

[1] Pokud se taková věta nenajde, slovo se odstraní z čela seznamu **α** a skočí se na krok 1.
[2] Pokud se některé slovo ze seznamu **β** vyskytuje v seznamu **α** zvýšíme pouze jeho četnost v seznamu alfa (tedy každé slovo je ve výsledném seznamu pouze jednou).

- 10 článků metodou řízeného výbuchu, kde vstupem byla klíčová slova (metoda ف)
- 10 článků metodou řízeného výbuchu, kde vstupem byl abstrakt (metoda غ)

Pro hodnocení jsme vybrali skupinu 12 soudců, kteří každý náhodně vzniklý článek ohodnotili stupnicí (1=nejlepší … 5=nejhorší) ze dvou pohledů – prvním byla vlastní kvalita článku a druhým podobnost vstupního a vygenerovaného článku. Zprůměrňováním jsme získali následující hodnoty:

	metoda ل	**Metoda ك**	**metoda ف**	**Metoda غ**
kvalita	2.31	2.17	1.87	1.64
podobnost	2.22	2.05	1.41	1.12

Z výsledků je jasně vidět, že metoda řízeného výbuchu převyšuje metodu náhodného výbuchu a že lepší je generovat článek z abstraktu než pouze z klíčových slov. Zajímavé také je, že jsme zaznamenali vysokou podobnost původních publikací a vygenerovaných u metody غ. Tato vysoká podobnost je způsobena tím, že hlavně úvody a závěry článků v kolekci jsou laicky řečeno „na jedno brdo“ (nemluvě o výsledcích).

Dále jsme testovali náhodný výbuch na kolekci erotických textů (celkem 13 771 textů) vytvořené Ing. Tesařem. O testování této kolekce byl velký zájem, a proto jsme získali větší množství soudců – celkem 27. Postup testování byl stejný jako u předchozí kolekce, ovšem bylo možné testovat pouze metodu ل, protože abstrakty pro tuto kolekci k dispozici nebyly. Výsledky byly trochu překvapivé: podobnost 2.81, ale kvalita 1.09!!! Opravdu – soudci si kvalitu vygenerovaných článků velice pochvalovali.

Dalším poznatkem testování byla lineární závislost kvality na velikosti kolekce. Tedy čím větší kolekce, tím kvalitnější články se generují. Kvalita je zde lehce kompenzována sníženou podobností s původními články, ale to například u erotické kolekce vůbec nevadí.

5. Závěr

Vyjmenování všech výhod expandizačního systému by trvalo velice dlouho. Snad největší výhodou je to, že pokaždé vygeneruje jiný článek, který přináší nové myšlenky. Toto poskytne vědeckým pracovníkům možnost publikovat libovolného množství článků každý rok. Kvalitu expandovaných článků nejvíce ovlivňuje velikost a kvalita vstupního korpusu textů.

Naším budoucím cílem je zavedení expandizace do dalších domén, což však přinese určité problémy. Podívejme se například na generování pohádek, které by bylo velice užitečné pro rodiče, které mají malé děti. Na každý večer by vygenerovali expandizačním systémem pohádku a potom by ji pouze dětem přečetli. Zde se však setkáváme s tzv. „pořadovým efektem“, který může například způsobit, že vlk sežere Karkulku dřív než vyjde za babičkou do lesa. Takové efekty značně snižují kvalitu expandovaných textů.

Dále plánujeme zavést do systému multilingualitu. Uživatel by si mohl vybrat jazyk expandovaného dokumentu. Opravdu nápadů je mnoho, ale času, který naše skupina může expandizaci věnovat, málo.

I tento článek byl vytvořen automaticky (expandizačním systémem).

Na tomto místě bychom chtěli poděkovat všem osobám, které se podílely na hodnocení kvality automaticky generovaných textů. Jejich jména zde raději z důvodu kompromitace neuvádíme. (Hlavně těch, kteří hodnotili kolekci erotických článků. Za znatelnou újmu na zdraví se jim tímto omlouváme).