

Techniki i eksperymenty dotyczące automatycznego streszczania tekstów artykułów prasowych.

Adam Dudczak Jerzy Stefanowski Dawid Weiss

Institute of Computing Science Poznan University of Technology

2 grudnia 2007

Plan prezentacji

- 1 Kilka pytań
- 2 Metody automatycznej selekcji zdań
 - Pozycja zdania w tekście
 - Automatyczne przydzielanie wagi słowom kluczowym
 - Łącuchy leksykalne
 - Automatyczne streszczanie dla języka polskiego
- 3 Lakon
- 4 Eksperymentalna ocena jakości

Co to jest streszczenie?

Co to jest streszczenie?

„Treść czegoś, ujęta krótko, zwięźle”

Uniwersalny słownik języka polskiego PWN, 2007

Czy tworzenie streszczeń jest trudne?

Na czym polega streszczenie?

Na czym polega streszczanie?

- Wybór najważniejszych informacji
- Zredagowanie nowej spójnej wypowiedzi

Jak ocenić jakość streszczenia?

Jak ocenić jakość streszczenia?

Taka ocena nie jest prosta:

- Jak streszczenie przekazuje to o czym była mowa w oryginale?
- Czy streszczenie zawiera błędy? np. ortograficzne
- Czy autor streszczenia nie dodał czegoś „od siebie”?
- Czy zawarte w streszczeniu informacje były rzeczywiście najważniejsze?

Czy trzeba rozumieć tekst żeby zrobić streszczenie?

Czy trzeba rozumieć tekst żeby zrobić streszczenie?

- Wiedza o budowie tekstu
- Statystyczna analiza tekstu

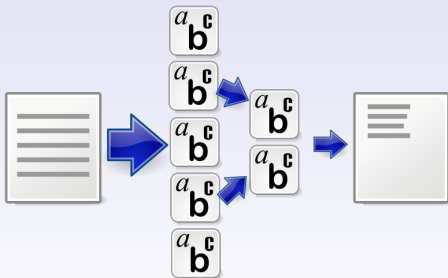
Czy komputer może wygenerować streszczenie?

Czy komputer może wygenerować streszczenie?

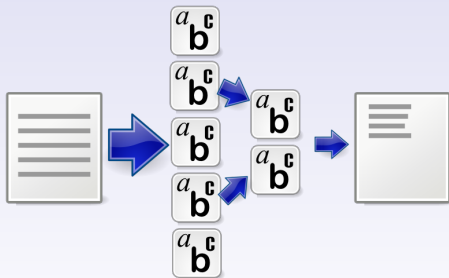
„The auto-abstract is perhaps the first example of a machine generated equivalent of completely intellectual task in the field of literature evaluation [Luhn, 1958].”

Na czym polega automatyczne streszczanie tekstu?

Na czym polega automatyczne streszczanie tekstu?



Na czym polega automatyczne streszczanie tekstu?



Ogólny podział:

- metody generatywne,
- metody oparte o selekcje pewnych fragmentów tekstu.

Metody generatywne

- Często wykorzystują pełną analizę syntaktyczną i semantyczną.
- Wybór najważniejszych informacji w oparciu model relacji między obiektami opisanymi w tekście.
- Synteza wynikowego streszczenia bądź. . .
- Wypełnianie pewnych szablonów np. raportów o porawaniach, raportów medycznych.
- Wymagające obliczeniowo, trudne w realizacji.
- Przykładowe systemy: Sumatra [Lie, 1998]

Metody oparte na selekcji

- Wybór najważniejszych słów, zdań, akapitów
- Bardzo różnorodne podejścia:
 - statystyczne
 - pogłębione - wykorzystanie dodatkowych źródeł wiedzy
 - wykorzystujące analizę semantyczną i syntaktyczną
- Niski koszt obliczeniowy (dla dwóch pierwszych)
- Możliwe do zastosowanie w rzeczywistych systemach

- Podobny sposób działania wszystkich tego typu metod
- Dwa ważne składniki algorytmu:
 - zbiór właściwości zdania, które decydują o jego wadze,
 - funkcja oceny istotności zdań

Najczęściej stosowane własności zdania

- Częstość występowania wyrazów w tekście
- Pozycja zdania w tekście
- Obecność pewnych zwrotów np. „Podsumowując. . .”
- Miary oceny ważności słów
- Długość zdania
- Obecność w zdaniu słów z tytułu lub rozpoczynających się od wielkiej litery.

Metoda wykorzystujące informacje o pozycji zdania

- Struktura artykułu prasowego, naukowego...
- Metoda jak najbardziej statystyczna
- Funkcja oceny:
 - najwyższe oceny: **zдания на początku акапитów**
 - najniższe: zdania na końcu akapitów
- Do **n** zdaniowego streszczenia wybierz **n** najlepiej ocenianych zdań (przy zachowaniu ich pierwotnej kolejności)

Dobre regulacje

Rynek kapitałowy w Polsce nie jest przeregulowany, natomiast zbyt mało jest spółek notowanych na giełdzie. Polski rynek prezentuje się dobrze, jest właściwie przygotowany pod względem organizacyjnym i prawnym.

Zdaniem prezesa giełdy warunki handlu na rynku giełdowym i pozagiełdowym będą inne, nie będzie więc między nimi konkurencji. Giełda jest przygotowana do rozwoju drugiego parkietu i stworzenia trzeciego.

- Dwa zdania: „Rynek kapitałowy...” i „Zdaniem prezesa...”

Dobre regulacje

Rynek kapitałowy w Polsce nie jest przeregulowany, natomiast zbyt mało jest spółek notowanych na giełdzie. Polski rynek prezentuje się dobrze, jest właściwie przygotowany pod względem organizacyjnym i prawnym.

Zdaniem prezesa giełdy warunki handlu na rynku giełdowym i pozagiełdowym będą inne, nie będzie więc między nimi konkurencji. Giełda jest przygotowana do rozwoju drugiego parkietu i stworzenia trzeciego.

- Dwa zdania: „Rynek kapitałowy...” i „Zdaniem prezesa...”
- Trzy zdania : „Rynek kapitałowy...”, „Polski rynek prezentuje...” oraz „Zdaniem prezesa...”

Term Frequency - Inverse Document Frequency

$$\text{tf-idf} = \text{tf} \cdot \log \left(\frac{|D|}{|d_j \supset t_j|} \right) \quad (1)$$

Gdzie:

- $|D|$ oznaczamy liczbę dokumentów w kolekcji,
- $|d_j \supset t_j|$ to liczba dokumentów (należących do kolekcji) które zawierają wystąpienia słowa kluczowego t_j .

$$\text{tf} = \frac{n_i}{\sum_k n_k} \quad (2)$$

- n_i liczba wystąpień danego wyrazu,
- $\sum_k n_k$ oznacza liczbę wszystkich słów

$$\text{bm25}(D, t) = \text{idf}(t) \cdot \frac{f(t, D) \cdot (k_1 + 1)}{f(t, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avdl}})}, \quad (3)$$

Gdzie:

- $f(t, D)$ – liczba wystąpień wyrazu t w dokumencie D .
- $|D|$ – długość dokumentu D
- avdl – średnia długość dokumentu w kolekcji
- k_1 i b – wartości stałe (przeważnie przyjmuje się $k_1 = 1,2$ i $b = 0,75$).

$$\text{IDF}(t) = \log \frac{N - n(t) + 0,5}{n(t) + 0,5}, \quad (4)$$

- N – liczba wszystkich dokumentów w kolekcji,
- $n(t)$ – liczba dokumentów w kolekcji zawierających słowo t .

Metody wykorzystujące schemat tf-idf i Okapi bm25

- Proste i sprawdzone
- Wymagają dostępu do referencyjnego zbioru dokumentów
 - 300 tysięcy artykułów z polskiej Wikipedii
- Dlaczego nie częstość występowania słów?
- Waga zdania to suma wag jego wyrazów
- Bardzo ważny jest dobór odpowiednich wartości minimalnych

Streszczanie przy użyciu tf-idf i Okapi bm25

Przebieg działania algorytmu:

- 1 Generowanie zbioru słów znaczących
 - 1 Wybór słów – czy tylko rzeczowniki czy wszystkie?
 - 2 Tworzenie rankingu słów,
- 2 Obliczanie wag zdań – polega na sumowaniu wag słów znaczących wchodzących w skład danego zdania,
- 3 Wybór zdań – do streszczenia wybierane są zdania z najwyższymi wagami.

Streszczenie przy użyciu tf-idf i Okapi bm25 - przykład

Rynek kapitałowy w Polsce nie jest przeregulowany, natomiast zbyt mało jest spółek, notowanych na giełdzie. Polski rynek prezentuje się dobrze, jest właściwie przygotowany pod względem organizacyjnym i prawnym.

słowo	tf-idf	bm25
rynek	0,331	1,061
giełdzie	0,237	0,801
spółek	0,2	0,677

Wady metod statystycznych

- Sens artykułu wyrażany jest przez kilka wyrazów o podobnym (synonimy) bądź węższym znaczeniu (hiponimy).
- Metody statystyczne nie są w stanie zidentyfikować tego typu zjawisk.
- Aby rozpoznawać synonimy, konieczna jest określenie w jakim znaczeniu dany wyraz występuje.
- Konieczne jest dostarczenie dodatkowej wiedzy w postaci tezauryusa
- Tezaurus **synonimy.ux.pl** pod redakcją p. Marcina Miłkowskiego

Spójność leksykalna

- [Barzilay and Elhadad, 1997] wprowadzają pojęcie spójności leksykalnej
- Spójność leksykalna to relacja która może zachodzić między dwoma wyrazami.
- Mamy dwa rodzaje spójności leksykalnej:
 - Reiteracja – zachodzi między wyrazem, a jego potwórzeńiem oraz między wyrazem a jego synonimem, bądź hiponimem.
 - Kolokacja – zachodzi gdy dwa wyrazy występują zawsze w podobnym kontekście znaczeniowym np. „Ona pracuje jako **nauczyciel** w **szkole**.”
- Spójność leksykalna może zachodzić również między wieloma wyrazami tworząc **łańcuchy leksykalne**.
- Reiteracje są dość łatwe do wykrycia.

Łańcuchy leksykalne - przykład

Zamek króla Korybuta był ogromny. O twierdzy tej słyszała chyba cała Europa. Mury cytadeli były bardzo grube i wydawały się sięgać nieba.

- Zidentyfikowane reiteracje:
 - powtórzenia : brak
 - synonimy : zamek, twierdza, cytadela
- Łańcuch leksykalny: zamek, twierdza, cytadela

Łańcuchy leksykalne - dezambiguacja

Zamek króla Korybuta był ogromny. O twierdzy tej słyszała chyba cała Europa. Mury cytadeli były bardzo grube i wydawały się sięgać nieba.

- Możliwe interpretacje słowa **zamek**:
 - Pierwsza grupa znaczeniowa: zamek, twierdza, cytadela
 - Druga grupa znaczeniowa : zamek, zasuwa, zatrask
- Wystąpienie słów twierdza i cytadela sugeruje, że słowo zostało użyte w kontekście pierwszej grupy znaczeniowej.

Budowanie łańcuchów leksykalnych

- Rozpatrywane są wszystkie możliwe interpretacje znaczenia wyrazów.
- Po osiągnięciu granicy akapitu usuwane są interpretacje, które nie znalazły żadnego potwierdzenia.
- Tworzenie łańcuchów (rozszerzanie istniejących) odbywa się po osiągnięciu granicy akapitu
- Dla gotowych łańcuchów obliczana jest wartość funkcji oceny.

Funkcja oceny łańcuchów leksykalnych

[Barzilay and Elhadad, 1997] zaproponowali następującą funkcję oceny:

$$\text{Ocena}(\text{Łańcuch}) = l \cdot h, \quad (5)$$

Gdzie:

- l – długość (ang. *length*) to liczba wystąpień elementów składowych łańcucha w tekście,
- h – jednorodność (ang. *homogeneity*) jest równe : (przez d oznaczamy liczbę elementów w łańcuchu):

$$h = 1 - \left(\frac{d}{l} \right). \quad (6)$$

Funkcja oceny łańcuchów leksykalnych

- Rozważmy dwa łańcuchy leksykalne (w nawiasach liczba wystąpień):
 - zamek (budowla, 1), cytadela (budowla, 1), twierdza (budowla, 1),
 - zamek (zasuwa, 1)
- Zgodnie z zaproponowaną przez [Barzilay and Elhadad, 1997] miarą jakości oba te łańcuchy mają taką samą siłę.
- Dłuższe łańcuchy są bardziej wartościowe w kontekście dezambiguacji wyrazów.

Nasza propozycja funkcji oceny to:

$$\text{Ocena}(\text{Łańcuch}) = l \cdot h + d \quad (7)$$

Gdzie d to liczba różnych leksemów z których dany łańcuch się składa.

Wybór najważniejszych zdań - przykład

Zamek króla Korybuta był ogromny, o władcy i jego twierdzy słyszała cała Europa. Mówiono o nim – wielki budowniczy.

- Na podstawie powyższego tekstu otrzymano następujące łańcuchy leksykalne (w nawiasach liczba wystąpień i numer zdania):
 - ① zamek (1,1), twierdza (1,1)
 - ② król(1,1), władca(1,1)
- Ocena tych łańcuchów jest równa 2.
- Waga danego zdania jest równa sumie ocen łańcuchów, których elementy znajdują się w danym zdaniu.

Wybór najważniejszych zdań - przykład

Zamek króla Korybuta był ogromny, o władcy i jego twierdzy słyszała cała Europa. Mówiono o nim – wielki budowniczy.

- Na podstawie powyższego tekstu otrzymano następujące łańcuchy leksykalne (w nawiasach liczba wystąpień i numer zdania):
 - ❶ zamek (1,1), twierdza (1,1)
 - ❷ król(1,1), władca(1,1)
- Ocena tych łańcuchów jest równa 2.
- Waga danego zdania jest równa sumie ocen łańcuchów, których elementy znajdują się w danym zdaniu.
- Łączna waga pierwszego zdania to: $2 \cdot 2 + 2 \cdot 2 = 8$.
- Zdania których wyrazy nie należą do żadnego z łańcuchów otrzymują wagę 0.

Wybór najważniejszych zdań - przykład

Zamek króla Korybuta był ogromny, o władcy i jego twierdzy słyszała cała Europa. Mówiono o nim – wielki budowniczy.

- Na podstawie powyższego tekstu otrzymano następujące łańcuchy leksykalne (w nawiasach liczba wystąpień i numer zdania):
 - ❶ zamek (1,1), twierdza (1,1)
 - ❷ król(1,1), władca(1,1)
- Ocena tych łańcuchów jest równa 2.
- Waga danego zdania jest równa sumie ocen łańcuchów, których elementy znajdują się w danym zdaniu.
- Łączna waga pierwszego zdania to: $2 \cdot 2 + 2 \cdot 2 = 8$.
- Zdania których wyrazy nie należą do żadnego z łańcuchów otrzymują wagę 0.
- Synteza streszczenia polega więc na wyborze n zdań o najwyższych wagach.

Automatyczne streszczanie dla języka polskiego

- Niewiele prac na temat generowanie automatycznych streszczeń dla języka polskiego.
 - System Polsumm2 - Politechnika Śląska
- W języku polskim duże zróżnicowanie form fleksyjnych wyrazów.
- Konieczna faza wstępnego przetwarzania tekstu.

Wstępne przetwarzanie języka polskiego

W kontekście streszczania ważne są następujące postulaty.

- Zachowanie pierwotnej struktury tekstu: podział na zdania i akapity, wyodrębnione nagłówki i tytuł tekstu.
- Pozyskanie informacji o cechach morfologicznych poszczególnych wyrazów i ich formie podstawowej.
- Rozpoznanie wyrazów pospolitych.

- Lakon od „mówić lakonicznie” czyli krótko, zwięźle i na temat.
- Projekt zawiera implementacje omówionych metod.
- Moduł pozwalający na elastyczną konfigurację wstępnego przetwarzania języka polskiego.
- Użytkownik może eksperymentować z zaimplementowanymi metodami przy wykorzystaniu prostej aplikacji
- Zawiera również szereg automatycznych analiz związanych z oceną jakości otrzymywanych streszczeń

Krótki pokaz...

Cele eksperymentu

Cele eksperymentu

- Pozyskać dane umożliwiające ocenę jakości zaimplementowanych metod streszczenia

Cele eksperymentu

Cele eksperymentu

- Pozyskać dane umożliwiające ocenę jakości zaimplementowanych metod streszczenia
- Czy w tekstach prasowych najważniejsze informacje znajdują się na początku?

Cele eksperymentu

Cele eksperymentu

- Pozyskać dane umożliwiające ocenę jakości zaimplementowanych metod streszczenia
- Czy w tekstach prasowych najważniejsze informacje znajdują się na początku?
- Jak bardzo będą się różnić między sobą streszczenia stworzone przez uczestników eksperymentu?

Założenia eksperymentu

- Zadanie uczestników: **wybór n najważniejszych zdań z prezentowanego tekstu.**
- Uczestnicy nie tworzą nowych wypowiedzi.
- Szybkie tworzenie streszczeń.
- Proste porównywanie z wynikami działania metod automatycznych.
- Długości streszczeń, które mają tworzyć uczestnicy to ok. 20% oryginalnego tekstu.

Korpus tekstów użyty w eksperymencie

- Dla celów eksperymentu pozyskano 10 artykułów prasowych.
- Zróżnicowana tematyka i długość
- Artykuły pochodzą z korpusu „Rzeczpospolitej” [Rzeczpospolita,] oraz z „Gazety Wyborczej” .

Przebieg eksperymentu

- Eksperyment trwał około miesiąca – od 13.05.2007 do 14.06.2007
- Aktywny udział wzięło 60 ochotników
- Stworzyli oni w sumie 285 streszczeń.
- Poszczególne artykuły posiadają od 27 do 30 streszczeń

Uczestnicy eksperymentu

- W eksperymencie wzięli udział głównie studenci oraz absolwenci studiów magisterskich.
- Osoby te w większości są (lub były) związane z uczelniami poznańskimi.
- Różnorodne specjalności : informatyka, ekonomia, farmacja, psychologia, prawo, filologia klasyczna, sinologia...

	ściśle	medyczne	humanistyczne	inne
liczba uczestników	34	3	21	2

Zebrane dane

Bibliografia



Barzilay, R. and Elhadad, M. (1997).
Using lexical chains for text summarization.
Intelligent Scalable Text Summarization Workshop (ISTS'97), pages 10–17.



Lie, D. (1998).
Sumatra: A system for automatic summary generation.
[on-line] <http://www.carp-technologies.nl/download/papers/SumatraTWLT14paper/SumatraTWLT14.html>.



Luhn, H. P. (1958).
The automatic creation of literature abstracts.
IBM Journal of Research and Development, pages 159–165.



Rzeczpospolita.
Korpus rzeczpospolitej.
[on-line] <http://www.cs.put.poznan.pl/dweiss/rzeczpospolita>.