Raport Badawczy

Research Report

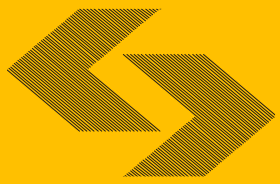
RB/58/2010

Przegląd wybranych technik segmentacji dokumentów tekstowych

M. Gajewski

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences



PRZEGLĄD WYBRANYCH TECHNIK SEGMENTACJI DOKUMENTÓW TEKSTOWYCH

Marek Gajewski
Studia Doktoranckie IBS PAN

W artykule rozważa się wybrane, znane z literatury techniki segmentacji stosowane przy przetwarzaniu dokumentów tekstowych. Omawia się pojęcie segmentu i różne możliwe zastosowania segmentacji. W szczególności, przedstawia się zarys koncepcji zastosowania segmentacji do wspomagania przetwarzania dokumentów tekstowych na użytek archiwisty.

***Słowa kluczowe**: segmentacja tekstu, wyszukiwanie informacji tekstowej, kategoryzacja dokumentów*

Wstęp

Przetwarzanie dokumentów tekstowych jest jednym z ważniejszych, nowoczesnych zastosowań informatyki. Obejmuje ono takie zadania jak: wyszukiwanie dokumentów spełniających potrzeby informacyjne użytkownika, wyrażone z użyciem zapytania; kategoryzacja dokumentów tekstowych, czy ekstrakcja informacji (ang. *information extraction*) z dokumentów tekstowych. Ludzie posługują się językiem naturalnym, dla przetwarzania którego nie stworzono dotychczas w pełni efektywnych narzędzi informatycznych. Zazwyczaj więc dokument, który ma być przetwarzany komputerowo musi być przedstawiony w pewnej uproszczonej reprezentacji interpretowalnej przez maszynę. Reprezentację dokumentu można na przykład zbudować zgodnie z tak zwanym modelem wektorowym (ang. *Vector Space Model).* W tym podejściu dokument reprezentowany jest za pomocą wektora, a poszczególne współrzędne tego wektora odpowiadają przyjętym *słowom kluczowym*. Zapytanie, określające jakich dokumentów użytkownik poszukuje, również przyjmuje w tym modelu postać wektora słów kluczowych. Dopasowanie dokumentu względem zapytania określa się na podstawie podobieństwa reprezentujących je wektorów. Faktycznie więc przyjmuje się, że wektory te reprezentują, odpowiednio, tematykę dokumentu i tematykę, która interesuje użytkownika. W przypadku zapytania ta tematyka jest zwykle dość wąsko określona i można ją względnie wiernie reprezentować z pomocą takiego wektora. W dokumencie jednak zazwyczaj poruszanych jest

wiele różnych wątków tematycznych i jego reprezentacja z użyciem jednego wektora może być mało adekwatna.

W literaturze często rozważany jest podział dokumentu na spójne tematycznie bądź funkcjonalnie fragmenty (segmenty). Callan [2] rozważa trzy typy takich fragmentów:

- wyróżnione jako logiczne elementy dokumentu, takie jak zdania czy akapity (ang. *discourse passages*),
- wyróżnione ze względu na ich jednolitą tematykę (ang. *semantic passages*), oraz
- wyróżnione „mechanicznie" poprzez podział dokumentu na fragmenty o ustalonej długości, wyrażonej na przykład liczbą wyrazów (ang. *window passages*).

W niniejszym tekście określenia „segment" będziemy używać w odniesieniu do drugiej z wyżej wymienionych klas.

Reprezentacja dokumentu jako zbioru wektorów reprezentujących poszczególne segmenty może przyczynić się do znacznego podniesienie efektywności realizacji różnych zadań z zakresu IR. Metody podziału treści dokumentu na spójne tematycznie fragmenty znane są pod pojęciem *segmentacji tekstu* (ang. *text segmentation)*, jak również jako *wykrywanie zmiany tematu* (ang. *topic change detection*) [4] [12] [14].

Zazwyczaj segmenty nie są jawnie wydzielone w treści dokumentu. Oznacza to konieczność zastosowania różnych zaawansowanych technik z zakresu IR, przetwarzania języka naturalnego (NLP) czy sztucznej inteligencji, celem ich wyodrębnienia.

Podział tekstu na segmenty znajduje wielorakie zastosowanie przy realizacji różnych zadań z zakresu wyszukiwania informacji tekstowej [12].

Po pierwsze, pozwala on na indeksowanie na poziomie segmentów zamiast na poziomie całego dokumentu. Przy założeniu, że wyodrębnione segmenty są spójne tematycznie powinno to umożliwić bardziej adekwatną ich reprezentację z użyciem np. metod modelu wektorowego. Jednocześnie, w odpowiedzi na zapytanie można użytkownikowi nie tylko przedstawić dokument, który powinien zaspokoić jego potrzeby informacyjne, lecz również można mu wskazać konkretne fragmenty tekstu najlepiej odpowiadające jego zapytaniu – będą to segmenty o największym stopniu dopasowania do zapytania.

Po drugie, wyróżnienie segmentów umożliwia automatyczne określenie tytułów i podtytułów sekcji tekstu, co może przyczynić się do podniesienia czytelności dużych objętościowo tekstów, oryginalnie pozbawionych struktury.

Po trzecie, segmentacja tekstu może być stosowana na potrzeby automatycznego podsumowywania tekstów.

Po czwarte, segmentacja może znaleźć zastosowanie przy automatycznej kategoryzacji (klasyfikacji) dokumentów tekstowych [9]. Dokument dzielony jest na segmenty, segmenty podlegają kategoryzacji, a następnie kategoria całego dokumentu określana jest na podstawie kategorii wyróżnionych w nim segmentów. Oczekuje się, że kategorie mogą być przypisane segmentom w sposób bardziej skuteczny, ze względu na zakładaną spójność tematyczną segmentów. Podejście to można rozszerzyć w ten sposób, żeby przy określaniu kategorii całego dokumentu uwzględniać nie tylko kategorie przypisane jego segmentom, ale również wzajemne położenie tych segmentów. W pewnych zastosowaniach może to mieć duże znaczenie. Ten obszar zastosowania segmentacji ma dla nas podstawowe znaczenie i będzie stanowił przedmiot naszych dalszych prac.

Pokrewnym zadaniem jest tzw. *wykrywanie i śledzenie wątku tematycznego* (ang. *topic detection and cracking, TDT*), w którym istotne jest wykrycie fragmentu(ów) tekstu dotyczącego określonego tematu, a segmentacja całego tekstu może być przy tym pomocnym narzędziem.

1. Przegląd wybranych metod segmentacji tekstu

W niniejszym punkcie szczegółowo opisujemy przykładowe metody segmentacji wybrane tak, aby zilustrować możliwie szerokie spektrum stosowanych rozwiązań.

W metodzie zaproponowanej przez Prince'a i Labadie [11] [12], przyjmuje się reprezentację dokumentów opartą na wstępnej analizie składniowej języka naturalnego. Zastosowany parser, specjalizowany dla języka francuskiego, określa części mowy poszczególnych wyrazów i fraz występujących w tekście (ang. *part-of-speech tagger, POS*). Dodatkowo, określa on również jaką rolę odgrywają one w zdaniu (podmiot, orzeczenie itd.). Dzięki temu, te same wyrazy występujące w różnych rolach mogą mieć przypisaną różną ważność dla reprezentacji dokumentu, co pozwala lepiej oddać znaczenie dokumentu.

Segment, rozumiany tu jako grupa zdań, reprezentowany jest jako wektor, którego postać określana jest w następujących krokach. Każdy wyraz i fraza, wyodrębnione we wstępnej analizie składniowej, reprezentowany jest jako wektor w 873 wymiarowej przestrzeni, przy czym każdy wymiar tej przestrzeni odpowiada jednemu pojęciu z ontologii zbudowanej na podstawie słownika Larousse'a. Zdanie reprezentowane jest jako liniowa kombinacja wspomnianych wcześniej wektorów reprezentujących zawarte w danym zdaniu wyrazy i frazy. Wagi w tej kombinacji liniowej określone są zależnie od części mowy

i zdania, jaką stanowi dany wyraz lub fraza, określonych jak to opisano w poprzednim akapicie. Segment reprezentowany jest przez centroid wektorów reprezentujących zdania składające się na ten segment, przy czym poszczególne zdania mają różny wpływ na określenie tego centroidu, zależnie od ich położenia względem początku segmentu: im dalej są one położone od początku tym mniejszy mają wpływ.

W opisywanej metodzie, przy wyodrębnianiu segmentów używana jest odległość pomiędzy dwoma zdaniami, która określona jest jako kąt pomiędzy wektorami reprezentującymi te zdania i wyrażona jest wzorem (1):

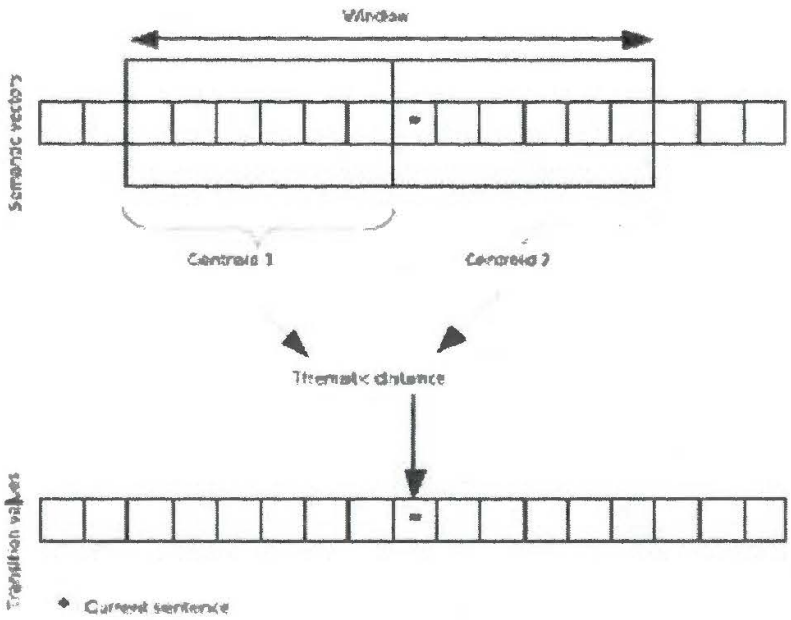
$$D_A(X,Y) = \arccos\frac{X \circ Y}{|X||Y|} \tag{1}$$

gdzie X i Y są wektorami reprezentującymi zdania, $\circ$ oznacza iloczyn skalarny wektorów, zaś $|X|$ oznacza normę euklidesową wektora X. Tak zdefiniowana odległość określona jest w [11] mianem *odległości tematycznej* (ang. *thematic distance*).

Sam algorytm segmentacji postępuje w następujący sposób. Przyjmuje się, że segmenty są jednorodne tematycznie i oddzielone są *strefami przejściowymi* (ang. *transition zones*), składającymi się ze zdań, które są odległe – w sensie odległości tematycznej – od zdań występujących w sąsiadujących segmentach. W celu wykrycia stref przejściowych używane jest *okno*, obejmujące ustaloną liczbę zdań i przesuwające się od początku do końca dokumentu. Przy każdym ustawieniu okna, dla centralnie w nim znajdującego się zdania z obliczana jest tzw. *wartość przejściowa* (ang. *transition value*). Wartość przejściowa jest obliczana jako odległość tematyczna (por. wzór (1)) pomiędzy wektorami reprezentującymi dwa segmenty. Pierwszy z tych segmentów obejmuje zdania położone na początku bieżącego okna aż do, ale z wyłączeniem, zdania z. Drugi segment obejmuje pozostałe zdania zawarte w bieżącym oknie, łącznie ze zdaniem z. Sposób obliczania wartości przejściowej zilustrowany jest na Rys. 1.

Jako strefa przejściowa uznawane jest zdanie lub grupa sąsiadujących zdań, dla których wartość przejściowa jest większa od pewnej, eksperymentalnie ustalonej, wartości progowej. Wskazanie w tekście stref przejściowych jest jednoznaczne z wyodrębnieniem segmentów. W kolejnych pracach [12] Prince i Labadie zaproponowali bardziej złożony sposób określania odległości pomiędzy zdaniami. Wprowadzona w tym celu *rozszerzona miara zgodności* stanowi kombinację odległości tematycznej określonej wzorem (1) oraz dodatkowych miar podobieństwa wektorów, uwzględniających wyłącznie te współrzędne

w nich występujące, które mają największe wartości. Autorzy w szeregu eksperymentów obliczeniowych potwierdzili lepsze wyniki uzyskane z użyciem takiej rozszerzonej miary zgodności.



Rysunek 1. Ilustracja sposobu obliczania wartości przejściowej w algorytmie Prince'a i Labadie (*źródło*: [11])

Hearst i Plaunt [8] opracowali algorytm *Text tilling* [8]. W tym algorytmie analizowany tekst dzielony jest na bloki o długości opisanej parametrem *k*, określającym liczbę zdań w bloku. Sugeruje się przyjąć dla *k* wartość równą średniej długości akapitów w dokumencie. Każdy blok traktowany jest jak samodzielny dokument i reprezentowany jest zgodnie z modelem wektorowym, przy zastosowaniu zmodyfikowanego schematu ważenia słów kluczowych *tf x IDF*. Następnie obliczane jest podobieństwo pomiędzy kolejnymi parami sąsiadujących bloków. Jest ono obliczane jako cosinus kąta pomiędzy wektorami reprezentującymi bloki. Otrzymuje się w ten sposób funkcję przypisującą każdemu blokowi wartość z przedziału [0,1] określającą jego podobieństwo do następnego bloku. Funkcja ta jest wygładzana w celu usunięcia lokalnych minimów. Granice pomiędzy segmentami wyznaczane są jako te bloki, dla których tak wygładzona funkcja uzyskuje minimum. W pracy [13] zaproponowano pewien wariant powyższej metody z użyciem mechanizmu Latent Semantic Indexing (LSI) do reprezentacji bloków.

Reynar [15] zaproponował algorytm segmentacji oparty na statystycznym modelowaniu języka naturalnego. Przyjmuje się w tym celu pewien model generowania tekstów w języku naturalnym. Następnie, podobnie jak w innych algorytmach, operuje się na parach sąsiadujących bloków tekstu, określanych tu mianem regionów. Dla każdej pary bloków sprawdza się czy prawdopodobieństwo wystąpienia drugiego z bloków (dokładniej: poszczególnych występujących w nim słów kluczowych) jest większe jeśli oblicza się je jako prawdopodobieństwo warunkowe względem wystąpienia pierwszego bloku czy też jeśli oblicza się je jako zdarzenie niezależne. Jeśli to drugie prawdopodobieństwo jest wyższe, to przyjmuje się, że wystąpiła zmiana tematu i miejsce to stanowi granicę pomiędzy segmentami.

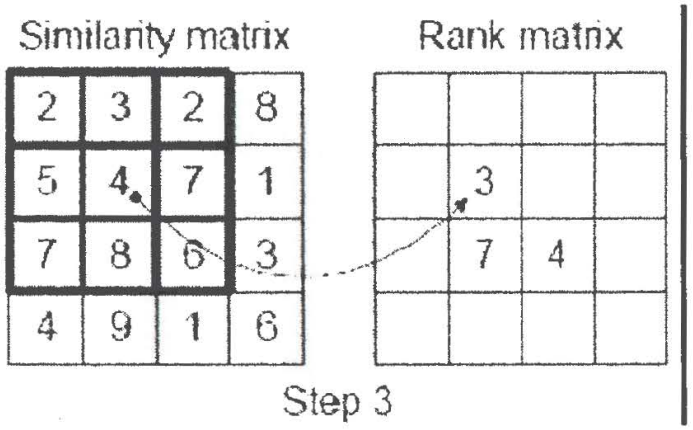
Reynar proponuje następnie zastosowanie wielu dodatkowych testów dla potwierdzenia czy wskazane w pierwszym kroku bloki stanowią faktycznie segmenty. Na przykład, proponuje on sprawdzenie czy w drugim bloku występuje wiele słów, które po raz pierwszy pojawiają się w analizowanym dokumencie (co traktuje się jako potwierdzenie istnienia granicy pomiędzy segmentami) lub czy w obrębie domniemanej granicy pomiędzy segmentami występują specyficzne dla tematyki dokumentu określenia, które mogą potwierdzać, że mamy do czynienia z odrębnymi tematycznie segmentami.

Algorytm C99 zaproponowany przez Choi [3] [4] charakteryzuje się zastosowaniem techniki podnoszenia kontrastu, zapożyczonej z dziedziny przetwarzania obrazów, oraz analizy skupień. Jest to powszechnie cytowany algorytm i poświęcimy jego opisowi nieco więcej miejsca.

Podstawową jednostką tekstu, która jest reprezentowana i przetwarzana w tym algorytmie jest zdanie. Poszczególne zdania są wstępnie przetwarzane w ten sposób, że usuwane są wyrazy nieznaczące (ang. *stopwords*) i pozostałe wyrazy zastępowane są przez ich rdzenie, w procesie określanym angielskim terminem *stemming*. Każde zdanie reprezentowane jest przez wektor określający częstość występowania w nim poszczególnych rdzeni. Następnie tworzona jest macierz bliskości zdań, przy czym bliskość dwóch zdań obliczana jest jako cosinus kąta pomiędzy reprezentującymi je wektorami (por. wzór (1), w którym arccos zastępujemy cos). Macierz ta jest oczywiście symetryczna i wzdłuż diagonali występują najwyższe wartości (równe 1), odpowiadające podobieństwu każdego zdania względem siebie samego.

Choi zauważa, że konkretne wartości występujące w macierzy podobieństwa zdań należy traktować raczej jakościowo niż ilościowo. Faktycznie, interpretacja konkretnych wartości cosinusa kąta pomiędzy wektorami reprezentującymi tak małe fragmenty tekstu, jakimi są zdania, w terminach absolutnych może być mylące. Co jest istotne, szczególnie z punktu widzenia segmentacji

tekstu, to stwierdzenie czy podobieństwo danych dwóch zdań jest duże czy małe w porównaniu do podobieństwa par zdań występujących w ich sąsiedztwie. W celu wychwycenia tego zjawiska stosuje się technikę zbliżoną do tej stosowanej w przetwarzaniu obrazów do ich wyostrzania (zwiększania kontrastu). Poszczególne elementy macierzy podobieństwa zdań zastępuje się ich *rangami*: liczbą sąsiednich elementów o mniejszej wartości. Stosuje się przy tym maskę w postaci podmacierzy o stosownie dobranych wymiarach. Na Rys. 2 przyjęto maskę 3 x 3, a w eksperymentach obliczeniowych opisanych w [3] stosowana maska ma wymiary 11 x 11. Otrzymaną w ten sposób macierz określa się mianem *macierzy rang* (ang. *rank matrix*).



Rysunek 2 Sposób przekształcania macierzy podobieństw w macierz rang (*źródło:* [3])

Segmenty wydzielane są z użyciem analizy skupień przeprowadzonej na zbiorze wszystkich zdań. Stosuje się przy tym algorytm rozdzielający (ang. *divisive clustering*), przy czym kryterium podziału skupień oparte jest na macierzy rang. Początkowo przyjmuje się, że wszystkie zdania tworzą jeden wielki segment. Załóżmy, że w danym kroku zbiór zdań podzielony jest już na kilka potencjalnych segmentów, $B=\{b_1,\ldots,b_m\}$. Do dalszego podziału wybierany jest jeden z segmentów b_i i takie w nim miejsce, że maksymalizowana jest tzw. *wewnętrzna gęstość* (ang. *inside density*) D nowego podziału otrzymanego z podziału B. Wewnętrzna gęstość obliczana jest wzorem (2):

$$D = \frac{\sum_{k=1}^{m} s_k}{\sum_{k=1}^{m} a_k} \tag{2}$$

przy czym s_k oznacza sumę rang zdań znajdujących się w segmencie b_k, zaś a_k oznacza liczbę par zdań w tym segmencie.

Optymalna liczba skupień określana jest w sposób automatyczny, następująco. Po każdym podziale oblicza się przyrost wewnętrznej zgodności: $D' - D$, gdzie D' oznacza nowy podział uzyskany z podziału D. Proces podziału kończy się wtedy, kiedy występuje duży spadek tego przyrostu. Oceniane jest to z użyciem wartości progowej określonej na podstawie średniej i wariancji z poprzednio zaobserwowanych przyrostów. Skupienia występujące w ostatnim podziale uzyskanym z użyciem algorytmu analizy skupień stanowią poszukiwane segmenty tekstu.

2. Zastosowanie segmentacji w złożonym zadaniu kategoryzacji

Segmentacja odgrywa znaczącą rolę w opracowywanej przez nas koncepcji systemu wspomagania pracy archiwisty. Zadanie polega na przypisaniu dokumentu tekstowego do odpowiedniej *grupy spraw*, a w ramach tej grupy do właściwych *akt sprawy*.

Automatyczne sporządzenie wykazu zbiorów dokumentów właściwych poszczególnym grupom spraw wydaje się być zadaniem nierealizowalnym. Należy zauważyć, iż dla większości grup spraw praktycznie niemożliwe jest sformułowanie warunków pozwalających na jednoznaczne określenie przynależności dokumentu do danej grupy rzeczowej. Z drugiej strony, pewne *typy dokumentów* mogą być częściej spotykane w grupie spraw A niż w grupie B. Z kolei, dla dokumentów poszczególnych typów powinno dać się określić charakterystyczną dla nich strukturę, w sensie składających się na nią segmentów.

W dalszej pracy nad zadaniem kategoryzacji musimy sprawdzić czy analizowany dokument należy do już założonej sprawy, alternatywnie czy jest to dokument inicjujący nową sprawę. Na tym etapie bardzo istotne jest poszukiwanie często trudnych do wychwycenia różnic pomiędzy pojedynczymi dokumentami. Tu ponownie, segmentacja może pomóc przy określaniu różnic/podobieństw dokumentów, ułatwiając ich skuteczną klasyfikacją do właściwych akt spraw.

Podsumowanie

W pracy omówiono ogólne zasady i wybrane algorytmy segmentacji dokumentów tekstowych. Nakreślono również koncepcję zastosowania segmentacji na użytek opracowywanego systemu wspomagania pracy archiwisty.

Literatura

[1] Amato F., Mazzeo A., Penta A., Picariello A. (2008): *Using NLP and Ontologies for Notary Document Management Systems.* 19th Intern. Workshop on Database and Expert Systems Applications (DEXA)

[2] Callan J. P. (1994): *Passage-Level Evidence in Document Retrieval.* Proc. of the 17th Annual Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval, Dublin, Ireland.

[3] Choi F. Y. Y (2000): *Advances in domain independent linear text segmentation.* Proc. of the 1st North American chapter of the Association for Computational Linguistics Conference, 26-33.

[4] Choi F. Y. Y. (2002): *Content-Based text navigation.* PhD thesis, Science and Engineering, University of Manchester

[5] Gajewski M. (2009): *Zastosowanie ontologii w zarządzaniu dokumentami elektronicznymi dla celów archiwizacji.* Analiza Systemowa w Finansach i Zarządzaniu. Wybrane problemy, Tom 11. Warszawa.

[6] Galley M., McKeown K. (2003*): Improving Word Sense Disambiguation in Lexical Chaining.* Proc. of the 18th Intern. Joint Conf. on Artificial Intelligence,. 1486-1488, Acapulco, Mexico.

[7] Hearst M. A. (1997): TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistic,* 23(1), 33-64,

[8] Hearst M. A., Plaunt Ch. (1993): Subtopic Structuring for Full-Length Document Access. Proc. of the Sixteenth Annual Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1993, 59-68.

[9] Kim J., Kim M. H (2004): An Evaluation of Passage-Based Text Categorization. *Journal of Intelligent Information Systems,* 23(1), 47-65.

[10] Moens M. F., Uyttendaele C. (1997): Automatic text structuring and categorization as a first step in summarizing legal cases. *Information Processing & Management,* 33 (6), 727-737.

[11] Prince V., Labadié A. (2007) Text segmentation based on document understanding for information retrieval. *Proc. of the 12th International Conf. on Applications of natural Language to Information Systems, Paris, France*

[12] Prince V., Labadié A. (2008): *Intended boundaries detection in topic change tracking for text segmentation.* 5th Intern.l Workshop on Natural Language Processing and Cognitive Science.

[13] Rehurek R. (2007*): On Dimensionality of Latent Semantic Indexing for Text Segmentation*, Proc. of the Intern. Multiconference on Computer Science and Information Technology, 347–356.

[14] Reynar J. C. (1998): *Topic Segmentation: Algorithms and applications.* PhD thesis, Computer and Information Science, University of Pennsylvania.

[15] Reynar J. C. (1999): *Statistical Models for Topic Segmentation.* Proc. of the 37th Annual Meeting of the Association for Computational Linguistics, 357-364, College Park, USA.