Segmentator regułowy do segmentacji raportów z akcji ratowniczo-gaśniczych PSP: metoda projektowania i ocena rozwiązania

STRESZCZENIE: W Państwowej Straży Pożarnej PSP do ewidencji zdarzeń służy system EWID-99 [1-3]. Po każdej interwencji PSP tworzony jest raport w wersji papierowej który przenoszony jest do wyżej wymienionego systemu elektronicznego [4]. Prelegent dokonał opracowania metody projektowania systemu informacyjnego bazującej na analizie zebranych tam danych tekstowych. Metoda ta opiera się na eksploracyjnej analizie tekstu do jego strukturalizacji [5, 6]. Prelegent wybrał do zaprojektowania system informacyjny na temat sieci hydrantów dla krajowego systemu ratowniczo-gaśniczego. Jeden z etapów proponowanej metody polegał na segmentacji raportów, który pociągał za sobą szereg problemów do rozwiązania omówionych podczas prelekcji.

Segmenty w literaturze poświęconej lingwistyce komputerowej i przetwarzaniu tekstów w języku naturalnym (*ang. natural language processing*), określa się też jako tokeny (*ang. tokens*). Podział ten polega na rozpoznawaniu granic między podstawowymi elementami tekstu w postaci segmentów. Segmentacja tekstu definiowana jest też jako liniowy podział tekstu na co najmniej dwóch poziomach [7]. Pierwszy poziom stanowi podział tekstu na jednostki, zwykle zdania, które mogą być przetwarzane składniowo niezależnie od innych jednostek tego samego poziomu. Drugi poziom stanowi segmentacja tekstu, prowadząca do tego, że tekst dzielony jest na jednostki, nazwane tokenami lub segmentami, którym przypisuje się interpretacje morfosyntaktyczne, czyli informacje o częściach mowy (rzeczownik, czasownik itp.) i wartościach odpowiednich kategorii morfosyntaktycznych (rodzaju, przypadku itp.). Zazwyczaj segmentacja w tym sensie nazywana jest tokenizacją. Dodatkowo dla poprawy dalszej interpretacji tekstu, a więc i jakości, ważne jest rozpoznawanie segmentów charakterystycznych dla tekstów danego typu, np.: dat, adresów, nazw ulic [8]. W badaniach prowadzonych przez prelegenta ważny aspekt stanowił podział tekstu na pierwszym poziomie. Ważne jest to z tego względu, że każdemu wydzielonemu segmentowi z raportu w procesie klasyfikacji nadawane jest znaczenie, określany jest jego kontekst. Odbywa się to poprzez analizę jego elementów składowych – wyrażeń. Na ich podstawie budowany jest klasyfikator który przydziela segment do jednej z wydzielonych klas semantycznych (określających kontekst). Nieprawidłowa segmentacja może doprowadzić więc nie tylko do niepoprawnego podziału zdania na części, ale także może doprowadzić do nieprawidłowej interpretacji semantycznej segmentu.

W literaturze dziedzinowej dotyczącej przetwarzania tekstów mało miejsca poświęca się metodom segmentacji na poziomie zdań i ich ocenom [7, 9-11]. W wystąpieniu omówiona zostanie opracowana przez prelegenta metoda podziału tekstu na segmenty oraz dokonana zostanie jej oceny. Skonstruowana metoda segmentacji polega na rozpoznawaniu granicy zdań w dokumentach tekstowych opisujących interwencje dokonywane przez służby ratownicze PSP. Okazało się, że zadanie to nie jest banalne w przypadku próby segmentacji badanych raportów. Do jego rozwiązania prelegent zaproponował metodę opartą o regułowe dzielenie tekstu na segmenty. Do realizacji procesu segmentacji zaprojektował on, w oparciu o formalną analizę pojęć (*ang. formal concept analysis – FCA*) [12-14], bazę wiedzy zawierającą używane w dokumentacji skróty oraz bazę reguł określającą warunki segmentacji. Całość rozwiązania, w postaci proponowanej metody, została poddana ocenie w odniesieniu do dwóch dostępnych dla prelegenta segmentatorów. Pierwszy z nich wykorzystywał rozszerzone reguły segmentacji (*ang. eXchange rule segmentation – SRX*)

[15]. Drugi natomiast stanowił komponent wchodzący w skład otwartego pakietu do analizy języka naturalnego (*ang. open natural language processing toolkit – openNLP*) [16].

Literatura

[1] Abakus: System EWID99. [on-line] [dostęp: 1 maja 2009] Dostępny w Internecie: http://www.ewid.pl/?set=rozw_ewid&gr=roz.

[2] Abakus: System EWIDSTAT. [on-line] [dostęp: 1 maja 2009] Dostępny w Internecie: http://www.ewid.pl/?set=ewidstat&gr=prod.

[3] Strona firmy abakus. [on-line] [dostęp: 1 marca 2009] Dostępny w Internecie: http://www.ewid.pl/?set=main&gr=aba.

[4] Rozporządzenie Ministra Spraw Wewnętrznych i Administracji z dnia 29 grudnia 1999 r. w sprawie szczegółowych zasad organizacji krajowego systemu ratowniczo-gaśniczego. Dz.U.99.111.1311 § 34 pkt. 5 i 6.

[5] Mirończuk M. Przegląd oraz zastosowanie metod eksploracji danych tekstowych do przetwarzania raportów z akcji ratowniczo-gaśniczych (preprint). Zeszyty Naukowe SGSP, 2011.

[6] PWN. Strukturalizacja. [dostęp: 1 kwietnia 2011] Dostępny w Internecie: http://sjp.pwn.pl/slownik/2576375/strukturalizacja.

[7] Przepiórkowski A. Techniki dezambiguacji morfo syntaktycznej. Powierzchniowe przetwarzanie języka polskiego. Warszawa: Akademicka oficyna wydawnicza EXIT, 2008. s. 17-45.

[8] Mykowiecka A. Elementy tekstu – segmenty, słowa, zdania. Inżynieria lingwistyczna Komputerowe przetwarzanie tekstów w języku naturalnym. Warszawa: Wydawnictwo PJWSTK, 2007. s. 65-83.

[9] Mykowiecka A. Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym. Warszawa: PJWSTK, 2007.

[10] Moens M. F. Information Extraction: Algorithms and Prospects in a Retrieval Context (The Information Retrieval Series). Springer, 2006.

[11] Mooney R. J., Bunescu R. C. Mining Knowledge from Text Using Information Extraction. SIGKDD Explorations, No 7, 2005, s. 3-10.

[12] Hesse W., Tilley T. Formal Concept Analysis Used for Software Analysis and Modelling. In: Ganter B., Stumme G. and Wille R., editors. Formal Concept Analysis: Springer Berlin / Heidelberg, 2005. s. 259-282.

[13] Priss U. Formal concept analysis in information science. Annual Rev Info Sci \& Technol, No 40, 2006, s. 521-543.

[14] Wolff K. E. A first course in formal concept analysis. [dostęp: 22 grudnia 2009] Dostępny w Internecie: http://www.fbmn.fh-darmstadt.de/home/wolff/Publikationen/A_First_Course_in_Formal_Concept_Analysis.pdf.

[15] Miłkowski M., Lipski J. Using SRX Standard for Sentence Segmentation. In: Vetulani Z., editor. Human Language Technology Challenges for Computer Science and Linguistics: Springer Berlin / Heidelberg, 2011. p. 172-182.

[16] openNLP. [dostęp: 1 kwietnia 20011] Dostępny w Internecie: http://incubator.apache.org/opennlp/.

Projekt współfinansowany ze środków Europejskiego Funduszu Społecznego w ramach Programu Operacyjnego Kapitał Ludzki Działanie 8.2 Transfer wiedzy, Poddziałanie 8.2.2 Regionalne strategie innowacji, budżetu państwa oraz środków Samorządu Województwa Podlaskiego.

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY