POLITECHNIKA BIAŁOSTOCKA

Wydział Informatyki

AUTOREFERAT ROZPRAWY DOKTORSKIEJ

mgr inż. Marcin Michał Mirończuk

Analiza danych tekstowych w projektowaniu wybranego systemu informacyjnego na przykładzie analizy dokumentacji zdarzeń krajowego systemu ratowniczo-gaśniczego

Promotor
dr hab. inż. Tadeusz Maciak

Białystok 2012

Spis treści

1. Wstęp

Po każdej interwencji służb ratowniczych Państwowej Straży Pożarnej PSP Kierujący Działaniami Ratowniczymi KDR sporządza papierową dokumentację opisującą przebieg interwencji. Forma tej dokumentacji w postaci formularza *Informacje ze zdarzenia* regulowana jest przez rozporządzenie [1]. Formularz ten zawiera m.in. sekcję pt. *Dane opisowe do informacji ze zdarzenia*. W sekcji tej KDR opisuje różne aspekty podjętych działań ratowniczo-gaśniczych za pomocą języka naturalnego. Po wypełnieniu formularza papierowego tekst jest wprowadzany w formie elektronicznej do systemu ewidencji zdarzeń EWID [2, 3, 4]. Omawiana sekcja dokumentacji papierowej, podzielona jest na sześć podpunktów: *opis przebiegu działań ratowniczych (zagrożenia i utrudnienia, zużyty i uszkodzony sprzęt)*, *opis jednostek przybyłych na miejsce zdarzenia*, *opis tego co uległo zniszczeniu lub spaleniu*, *warunki atmosferyczne*, *wnioski i uwagi wynikające z przebiegu działań ratowniczych* oraz *inne uwagi dotyczące danych wypełnianych w formularzu odnośnie zdarzenia*. W systemie EWID brak jest podziału na takie podpunkty i zapisywany jest jednolity raport tekstowy wyrażony za pomocą języka naturalnego. W dalszej części opracowania pod pojęciem *tekst* należy rozumieć opisy wyrażone językiem naturalnym znajdujące się w elektronicznej sekcji *Dane opisowe do informacji ze zdarzenia* systemu ewidencji. W tekstach tych znajdują się ważne informacje oraz wiedza dziedzinowa na temat np. sposobu neutralizacji powstałych zagrożeń czy też rodzaju użytego sprzętu do ich likwidacji.

Autor podczas swoich badań wykazał brak możliwości zastosowania omawianych tekstów bezpośrednio do analizy. Wynika to m.in. z faktu, że w rezultacie przeszukiwania sekcji elektronicznej KDR może dostać nieoczekiwane rezultaty np. kierując zapytanie o *hydranty przy ulicy Mickiewicza* system może zwrócić informacje nie tylko o hydrantach ale także o wszystkich akcjach ratowniczo-gaśniczych przy tej ulicy [5]. Rozwiązanie tych problemów miała stanowić eksploracyjna analiza danych tekstowych (*ang. text mining - TM*) [6, 7, 8, 9] będąca specjalną odmianą (działającą na tekście) procesu odkrywania wiedzy w bazach danych (*ang. knowledge discovery in databases - KDD*) [10, 11, 12, 13, 14]. Aktualnie istnieją aplikacje do eksploracyjnej analizy danych tekstowych [15], które najczęściej współpracują z komponentami z zakresu przetwarzania języka naturalnego (*ang. natural language processing - NLP*) [16]. Zazwyczaj działanie tych pierwszych ogranicza się do analizy dokumentów tekstowych jako całości. Eksploracyjna analiza danych tekstowych pomija badanie zależności gramatycznych i morfologicznych na poziomie pojedynczych wyrażeń, które są domeną dziedziny przetwarzania języka naturalnego. Ewentualnie NLP stanowi uzupełnienie procesu wstępnego przetwarzania dokumentów tekstowych poprzez dostarczanie rozwiązań z zakresu np. lematyzacji, stemmingu [16]. Zarówno jednak pierwsze jak i drugie podejście jest niewystarczające z tego względu, że pomija badanie segmentu, części obszerniejszego tekstu jako samodzielnego obiektu, który może nieść sam w sobie informację. Segment w

kontekście badań stanowi element tekstu w postaci zdania, które ma określony początek oraz koniec. Zazwyczaj początek zdania rozpoczyna się od dużej litery i kończy się znakiem interpunkcyjnym w postaci *"."*, *"!"*, *"?"* etc. Badanie w takim kontekście segmentu, jak i nawet próba jego wyekstrahowania z dostępnych tekstów dziedzinowych, okazało się nietrywialne i znaczące w prowadzonych przez autora eksperymentach nad tekstem i jego strukturalizacją w celu zaprojektowania wybranego systemu informacyjnego.

Autoreferat składa się z czterech części, w których autor kolejno omawia elementy proponowanej przez siebie metody projektowania systemu informacyjnego oraz opisuje przeprowadzone badania w celu zademonstrowania jej realizacji. W punkcie 2 przedstawiono cel oraz zakres zrealizowanych badań jak i tezę pracy. Aspekty przeprowadzonej przez autora analizy jakościowej zostały opisane w punkcie 3. W punkcie 4 autor przedstawił projekty poszczególnych składowych proponowanego procesu do strukturalizacji danych tekstowych. Punkt 5 zawiera wnioski z przeprowadzonych badań i projektowania wybranego systemu informacyjnego na podstawie analizy (strukturalizacji) danych tekstowych w proponowanym autorskim procesie.

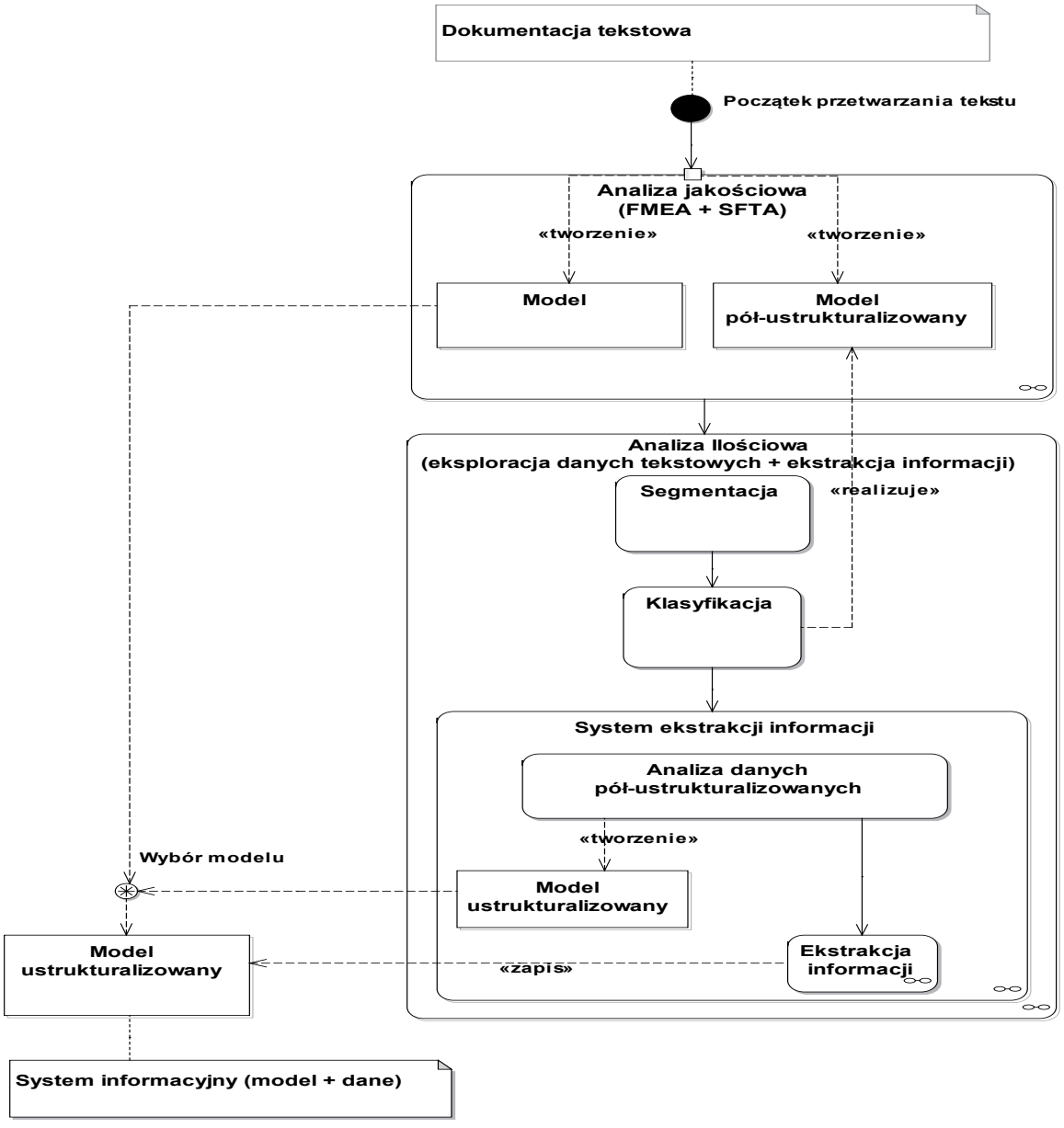
2. Cel i zakres zrealizowanych badań oraz teza pracy

Celem opisanego w pracy badania było opracowanie zintegrowanej metody do projektowania systemu informacyjnego SI stanowiącego narzędzie do realizacji procesów informacyjnych [17, 18] w oparciu o proces odkrywania wiedzy z baz danych tekstowych [19]. W pracy zaproponowano nazwę dla takiego procesu w postaci - projektowanie SI sterowane danymi tekstowymi (*ang. text driven software design*). Nazwa ta ma odróżniać i podkreślać specyficzny charakter przedsięwzięcia od tradycyjnego procesu eksploracji danych tekstowych i odkrywania wiedzy z baz danych, które uwydatniają aspekt związany z tym, iż wiedza jest końcowym produktem odkrywania sterowanego danymi (*ang. data-driven discovery*).

W pracy dokonano przedstawienia problematyki związanej z proponowaną przez autora ogólną metodą projektowania SI realizowaną za pomocą ww. procesu. Całościowo wykazano możliwość dostosowania tego procesu, opartego o eksploracyjną analizę danych, do strukturalizacji dokumentacji tekstowej, wyrażonej za pomocą języka naturalnego i projektowania za jego pomocą SI. Studium przypadku (*ang. case study*) realizacji skonstruowanej metody stanowiła analiza dokumentacji zdarzeń krajowego systemu ratowniczo-gaśniczego, która została pozyskana z jednostki ratowniczo-gaśniczej JRG. Istniał znaczny problem z jej pozyskaniem ze względu na ustawę o ochronie danych osobowych. Stąd też przed dostarczeniem jej autorowi poddana została ona żmudnemu i pracochłonnemu procesowi czyszczenia z danych osobowych. Ogranicza to już na wstępie możliwości przeprowadzenia pewnych analiz. W sumie pozyskano do badań 28800 raportów, z których

do dalszych analiz wybrano 3735 tekstów. Ponadto autor nie wiedział, z której dokładnie JRG pochodziły uzyskane teksty. Wszystkie te elementy wpływały ograniczająco na możliwości przeprowadzonej analizy.

Pozyskana dokumentacja tekstowa w postaci dokumentacji zdarzeń, w kontekście projektowania SI, stanowiła jego wstępną specyfikację i źródło danych operacyjnej bazy danych. Za pomocą zaproponowanej metody do jej analizy, zaprojektowano i uzupełniono danymi wybrany system informacyjny. Szczegółowe podstawy teoretyczno-praktyczne zaproponowanej i przebadanej metody do projektowania SI z wykorzystaniem analizy jakościowej i ilościowej, prezentuje rysunek 1.



Rysunek 1. Ogólny schemat blokowy zrealizowanych badań. Źródło: [opracowanie własne]

Rysunek 1 prezentuje dwa podstawowe bloki analizy jakościowej i ilościowej. Podczas pierwszej wymienionej analizy zostały zbadane teksty pod kątem występowania w nich określonych struktur w postaci scenariuszy (*ang. script*) i sieci semantycznych (*ang. semantic nets*) [20, 21]. Skonstruowano ogólną sieć pojęć znajdujących się w badanych tekstach. Następnie za pomocą skonstruowanej przez autora zmodyfikowanej analizy przyczyn i skutków błędów (*ang. failure modes and effects analysis - FMEA*) zawierającej drzewo analizy błędów oprogramowania (*ang. software failure tree analysis - SFTA*) zbadano teksty pod kątem wyszukiwania w nich określonej informacji. W pierwszym kroku opracowano szczegółową sieć semantyczną przykładowego systemu informacyjnego. W drugim kroku pojęcia z tej sieci zorganizowano w bardziej strukturalną reprezentację w postaci ramek (*ang. frames*), wyrażonych za pomocą interfejsu obiektowego [5]. Podczas analizy ilościowej, zaprojektowano trzy niezbędne elementy do proponowanej strukturalizacji tekstu. Elementy te stanowiły: segmentator, klasyfikator oraz system ekstrakcji informacji.

Końcowy ustrukturalizowany model był efektem analizy i synchronizacji dwóch modeli tj. pierwszego otrzymanego ze wstępnej, jakościowej analizy danych tekstowych i drugiego utworzonego podczas projektu i analiz systemu ekstrakcji informacji. Realizacja wszystkich ww. elementów analiz oraz ich weryfikacja poprzez implementację i eksperymenty była konieczna do zaprojektowania wybranego, przykładowego SI i tym samym zaproponowania oryginalnego procesu projektowania i wytwarzania SI na podstawie analizy tekstu.

W odróżnieniu do klasycznej ekstakcji informacji [22, 16], proponowana autorska metoda wprowadza analizę jakościową danych tekstowych, ich segmentację wraz z klasyfikacją segmentów do wyznaczonych w analizie jakościowej klas semantycznych oraz ich strukturalizację. Różnice dotyczące ekstrakcji informacji na poziomie segmentów w odniesieniu do klasycznej ekstrakcji informacji zostały przedstawione przez autora w pracy [23].

Teza rozprawy brzmi następująco:

Eksploracyjna analiza danych tekstowych może być zastosowana na etapie projektowania systemu informacyjnego stanowiącego narzędzie do realizacji procesów informacyjnych

W celu jej udowodnienia, posłużono się trzema tezami pomocniczymi:

1. Teza pomocnicza 1:
   Tekst występujący w sekcji Dane opisowe do informacji ze zdarzenia nie nadaje się wprost do wyszukiwania potrzebnej informacji
2. Teza pomocnicza 2:
   Na podstawie analizy składników budujących segment tekstu można rozpoznać jego semantykę w tekście. Dokonać tego można za pomocą procesu eksploracji danych tekstowych w postaci segmentów do ich klasyfikacji

3. Teza pomocnicza 3:
   Na podstawie wydzielonych i poklasyfikowanych segmentów tekstów można zamodelować i wyekstrahować wybrany rodzaj informacji

3. Analiza jakościowa danych tekstowych

Analiza jakościowa tekstów przebiegała dwustopniowo. Najpierw autor dokonał ich analizy pod kątem występowania w nich określonych struktur w postaci scenariuszy oraz sieci semantycznych. Scenariusze odpowiadały pół-ustrukturalizowanemu opisowi tekstu (model pół-ustrukturalizowany, pół-ustrukturalizowany tekst lub pół-ustrukturalizowany przypadek zdarzenia). Na podstawie analizy tekstów autor ustalił, że budujące go segmenty można przydzielić do pięciu klas semantycznych. Opracował także reguły do przeprowadzenia takiego przydziału. Na skonstruowane klasy semantyczne składały się takie kategorie jak: *operacje*, *ogólna*, *sprzęt*, *szkody* oraz *meteo*. W ten sposób otrzymano formę zbliżoną do wejściowej, papierowej reprezentacji sekcji *Dane opisowe do informacji ze zdarzenia*. Na przykładzie nieustrukturalizowanego tekstu 1 zademonstrowano sposób przeprowadzenia analizy.

Nieustrukturalizowany tekst 1

„Po dojechaniu na miejsce zdarzenia stwierdzono pożar instalacji elektrycznej w skrzynce z bezpiecznikami na klatce schodowej. Działania psp polegały na oddymieniu i przewietrzeniu klatki schodowej na parterze. Na miejsce zdarzenia przybyło pogotowie energetyczne z ul. chrzanowskiego celem zabezpieczenia instalacji. Sprawny hydrant nr 34922 ul. Szaserów 99."

W wyniku analizy nieustrukturalizowanego tekstu 1 tj. po przypisaniu segmentów, za pomocą reguł, do klas sematycznych otrzymywany jest pół-ustrukturalizowany model. Model ten prezentuje tabela 1.

Tabela 1. Pół-ustrukturalizowany użyteczny przypadek zdarzenia. Źródło: [opracowanie własne]

L.p.	Zdanie oryginalne	Klasa
1	Po dojechaniu na miejsce zdarzenia stwierdzono pożar instalacji elektrycznej w skrzynce z bezpiecznikami na klatce schodowej	Ogólna
2	Działania psp polegały na oddymieniu i przewietrzeniu klatki schodowej na parterze	Operacje
3	Na miejsce zdarzenia przybyło pogotowie energetyczne z ul. chrzanowskiego celem zabezpieczenia instalacji	Ogólna
4	Sprawny hydrant nr 34922 ul. Szaserów 99	Sprzęt

Tabela 1 prezentuje pół-ustrukuralizowany model, który powstał przy założeniu tzw. kompozycyjności semantyki, która została opisana dalej w referacie przy omówieniu badań ilościowych. Analizując dane zebrane w tabeli 1 widać, że tekst został rozbity na cztery

segmenty. Segmenty te na podstawie analizy budujących ich wyrażeń zaklasyfikowano odpowiednio do klas semantycznych.

Po skonstruowaniu scenariuszy czyli pół-ustrukturalizowanego modelu oraz sieci semantycznej autor przystąpił do analizy tych struktur oraz tekstu pod kątem wyszukiwania w nich informacji. W tym celu zaprojektował on zmodyfikowaną analizę FMEA z elementami SFTA [5]. Analiza ta w ogólnym przypadku może służyć do odnalezienia schematów opisów w dostępnej dokumentacji tekstowej czy też do potwierdzenia lub sfalsyfikowania możliwości wyszukania w nich zdefiniowanej informacji jak i zaproponowania bardziej odpowiedniej struktury danych. W opisywanym przypadku analizy dokumentacji zdarzeń, analiza ta została wykorzystana do potwierdzenia lub sfalsyfikowania możliwości:

— zastosowania aktualnego systemu ewidencji zdarzeń i przechowywanych w nim nieustrukturalizowanych danych tekstowych jako operacyjnej bazy danych,
— zastosowania przechowywanych w systemie ewidencji zdarzeń nieustrukturalizowanych danych tekstowych jako operacyjnej bazy danych, w przypadku zastosowania procesu ich strukturalizacji.

Całość proponowanej analizy wykazała, że dane zawarte w części opisowej do informacji o zdarzeniach nie nadają się do wyszukiwania potrzebnej informacji. Autor wysunął również na tym etapie propozycję procesu do strukturalizacji danych i strukturalnego ich zapisu. Działania te miały na celu zlikwidować mankamenty związane z przechowywaniem informacji wyrażonej za pomocą języka naturalnego.

4. Analiza ilościowa danych tekstowych

Etap analizy ilościowej zrealizowany podczas badań określił możliwość zastosowania procesu strukturalizacji dostępnych tekstów i zaprojektowania na jego podstawie systemu informacyjnego. Dzięki temu etapowi i przeprowadzonym w nim eksperymentom określono, zaproponowano i wybrano modele najlepiej nadające się do tego celu. Zaprojektowano i zaimplementowano także niezbędne oprogramowanie do jego zrealizowania. Proces strukturalizacji składał się z trzech podstawowych komponentów. Komponenty te można analizować jako trzy oddzielne systemy współdziałające ze sobą, do których należą: system regułowy do segmentacji (Segmentator Regułowy - SR) omówiony w podpunkcie 4.1, system klasyfikacji semantycznej segmentów (SKSS) omówiony w podpunkcie 4.2 oraz system ekstrakcji informacji (SEIt - litera t w akronimie określa temat), omówiony w podpunkcie 4.3.

Wymienione wyżej systemy można traktować jako systemy ekspertowe. Celem takich platform jest m.in. zautomatyzowanie pewnych aspektów działań ludzkich czy też w zupełności wyręczenie człowieka w pewnych zadaniach [21]. Proponowane dwa systemy w postaci

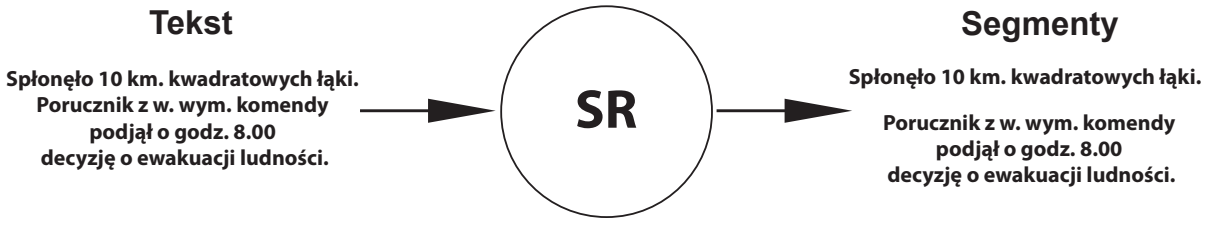
SR oraz SEI zrealizowane zostały w oparciu o metody i techniki z zakresu symbolicznej sztucznej inteligencji [20]. Ponadto stanowiły one odmianę systemów ekspertowych w postaci regułowych systemów ekspertowych [20, 21]. Ich działanie było oparte o reguły jak i płynące z nich wnioski. Do zaprojektowania tych rozwiązań tj. odpowiednich baz wiedzy reguł wykorzystano formalną analizę pojęć (*ang. formal concept analysis – FCA*) [24, 25, 26]. Dzięki niej uzyskano diagramy liniowe (*ang. lines diagrams*). Diagramy te stanowiły opis wiedzy deklaratywnej w postaci *"wiadomo, że..."*. W przypadku SR wiadomo było, że m.in. istnieją specjalistyczne skróty czy też skróty niepoprawne ale dające się zinterpretować np. skrót *sam. os.* oznaczał samochód osobowy. Interpretacja tych diagramów dostarczała natomiast wiedzy proceduralnej do projektowania i implementacji oprogramowania. Wiedza proceduralna wyraża się poprzez stwierdzenie *"wiem jak..."*. W przypadku SR wiadomo było jak m.in. skonstruować oprogramowanie poprawnie segmentujące tekst w oparciu o utworzone diagramy liniowe.

System ekstrakcji informacji semantycznej również był pewnego rodzaju systemem ekspertowym. W odróżnieniu od dwóch poprzednich platform, wykorzystywał on aspekty związane z obliczeniową sztuczną inteligencją [20]. Działanie jego w większości oparte było o metody i techniki z zakresu eksploracyjnej analizy danych tekstowych oraz przetwarzania języka naturalnego.

4.1. System segmentacji regułowej - Segmentator Regułowy (SR)

Segmentacja tekstu stanowi podział polegający na rozpoznawaniu granic między podstawowymi elementami tekstu w postaci segmentów. Podziału tekstu na jednostki, zwykle zdania, dokonuje się w celu ich przetwarzania składniowego, niezależnie od innych jednostek tego samego poziomu [16, 27]. Segmentacja dostępnego tekstu została przeprowadzono za pomocą autorskiego rozwiązania w postaci SR. Działanie jego zostało także porównane z innymi dwoma rozwiązaniami. Pierwsze z nich wykorzystywało rozszerzone reguły segmentacji (*ang. eXchange rules segmentation – SRX*) [28]. Drugi natomiast stanowił komponent wchodzący w skład otwartego pakietu do analizy języka naturalnego (*ang. open natural language processing toolkit – openNLP*) [29].

Ogólne działanie rozwiązania służącego do segmentacji tekstu przedstawia rysunek 2.



Rysunek 2. Ogólny proces segmentacji tekstu. Źródło: [opracowanie własne]

Rysunek 2 przedstawia ogólny proces segmentacji. Na rysunku tym widać tekst wejściowy, który w wyniku działania procesu segmentacji zaimplementowanego za pomocą np. proponowanych jak i badanych rozwiązań, podzielony został na dwa segmenty. Widać, że w przypadku zastosowania prostej reguły r_1*: każda kropka kończy segment*, można otrzymać aż sześć segmentów zamiast docelowych dwóch. Chcąc w dalszej kolejności analizować dostępne teksty autor nie mógł dopuścić do powstawania tak dużych i nieakceptowalnych błędów. W wyniku przeprowadzonej przez niego analizy ustalił, że należy zbudować odpowiednią bazę wiedzy o skrótach występujących w badanych tekstach oraz bazę reguł podziału tekstu na segmenty.

Posługując się utworzonymi bazami wiedzy zaproponowano oraz zaprojektowano oprogramowanie przetwarzające (segmentujące) raporty. Za pomocą skonstruowanego oprogramowania do segmentacji autor wykazał także, że można zaproponować odpowiednie rozwiązanie do tego celu. Uzyskał również za pomocą własnego rozwiązania znacznie lepsze rezultaty od dostępnych rozwiązań. W głównej mierze było to spowodowane dokładną analizą charakterystycznych skrótów występujących w tekstach. Stąd też płynął wniosek, że dla badanych dziedzinowych tekstów należy tworzyć rozwiązania dedykowane.

4.2. System klasyfikacji semantycznej segmentów (SKSS)

Opisany poniżej SKSS bazował faktycznie na założeniach teorii zależności pojęciowej (*ang. conceptual dependency theory*) opracowanej pod koniec lat 60 i semantyce kompozycyjnej. Wymieniona teoria bazuje na tezie, że to nie syntaktyka powinna być punktem wyjścia do analizy semantycznej, ale *pojęcia*, a ściślej mówiąc wzajemne zależności (relacje) między pojęciami [20]. Semantyka kompozycyjna zajmuje się budową znaczeń większych konstrukcji składniowych na podstawie znaczeń ich składników czyli słów i wyrażeń [27]. Teza pomocnicza nr. 2 przyjęta przez autora jest praktycznie zbieżna z tezą przedstawioną w ww. teorii. Różnica polega na tym, że autor rozprawy zdefiniował i wyznaczył konkretne narzędzie, które służyło do tego celu w postaci procesu do eksploracji danych tekstowych realizującego zadanie klasyfikacji. Poprzez klasyfikację wyznaczona została semantyka segmentu. Z drugiej strony o proponowanym przez autora procesie można mówić w kontekście ekstrakcji informacji semantycznej. Stąd druga, alternatywna nazwa *systemu ekstrakcji informacji semantycznej* SEIS.

Odnalezienie znaczenia segmentu w tekście odbywa się poprzez ustalenie a następnie nadanie mu etykiety semantycznej w postaci nazwy klasy. Semantyka, czyli znaczenie segmentu było określone poprzez wyrażenia, które budują segment. Znaczenie określane było poprzez funkcję opisującą połączenie poszczególnych wyrażeń w segmencie. Do określania znaczenia segmentu autor zaproponował zastosowanie funkcji, modelu klasyfikacji z zakresu sztucznej inteligencji. Ponadto proponowana analiza semantyczna była formą pośrednią między

klasyfikacją całych dokumentów tekstowych a badaniem poszczególnych wyrażeń. Do jej przeprowadzenia autor użył zbioru 3735 tekstów. Zostały one, za pomocą zaprojektowanego i zaimplementowanego przez autora segmentatora omówionego w podpunkcie 4.1, podzielone na segmenty. Otrzymany referencyjny zbiór 12753 segmentów manualnie zaetykietowano (segmenty przydzielono do klas). Na tak utworzonym zbiorze segmentów autor sprawdził działanie trzech rodzajów klasyfikatorów: k-najbliższych sąsiadów (*ang. k-nearest neighbor*), naiwnego Bayesa oraz centroidalnego (Rocchio) z różnymi autorskimi jego modyfikacjami [30, 31].

W rezultacie przeprowadzonych eksperymentów nad klasyfikacją segmentów, osiągnięto zadowalające rezultaty sięgające 90% poprawnie sklasyfikowanych segmentów. Ponadto otrzymane rezultaty z eksperymentów dokonywanych na korpusie polskich tekstów w postaci segmentów są zasadniczo zbieżne z rezultatami otrzymywanymi na świecie tj. z eksperymentami na tekstach angielskich. Proponowane natomiast przez autora odmiany klasyfikatora centroidalnego dawały polepszenie procesu klasyfikacji.

4.3. System ekstrakcji informacji na wybrany temat (SEIt)

Ekstrakcja informacji (*ang. information extraction*) jest to identyfikacja, polegająca na odnajdywaniu właściwej informacji w nieustrukturyzowanych danych tekstowych wyrażonych za pomocą języka naturalnego. Proces ten jest zgodny z klasyfikacją polegającą na strukturyzowaniu poprzez nadawanie klas semantycznych dla wybranych elementów tekstu. Proces ten czyni informację zawartą w tekście bardziej właściwą i przydatną w realizowanych zdaniach. W kontekście aktualnie przedstawianych badań autora, ekstrakcja informacji polegała na rozpoznawaniu nazw encji. Zadanie to bazuje na rozpoznawaniu i klasyfikowaniu wykrytych wyrażeń z tekstu takich jak: nazwy osób, firm, lokalizacji, dowódców, wozów bojowych etc.

Proponowany przez autora proces budowy wybranego modelu oraz ekstrakcji do niego informacji z raportów opisujących interwencje PSP zawiera dwa główne tory przetwarzania. Pierwszy tor związany jest z tworzeniem słowników wyrażeń budujących badane segmenty oraz modelu SI. Drugi tor natomiast dotyczy ekstrakcji informacji do utworzonego modelu. Słowniki jak i silnik ekstrakcji informacji w postaci stosu odpowiednio zaprojektowanych i ułożonych reguł zostały wykonane przy użyciu analizy opartej o formalną analizę pojęć.

W badaniach nad systemem ekstrakcji informacji na wybrany, przykładowy temat wyselekcjonowano 1416 segmentów opisujących *punkty czerpania wody – Hydranty*. Na podstawie ich autorskiej analizy w skonstruowanym i zaimplementowanym procesie ekstrakcji została pozyskana informacja o ich m.in. względnej lokalizacji czy też sprawności. Dysponując informacją o lokalizacji, autor w procesie geokodowania [32] ustalił też względną szerokość oraz długość geograficzną danego obiektu. Dzięki temu zabiegowi uzyskał on dodatkowy

efekt w postaci wizualizacji danych tj. położenia zarejestrowanych hydrantów dostępnych w zbudowanym rejestrze.

Autor w dodatku do rozprawy zaprezentował także drugi przykład zastosowania proponowanej metody projektowej SI oraz ekstrakcji informacji. Dodatkowe studium przypadku dotyczyło analizy raportów opisujących wypadki samochodowe. Do analizy pozyskano 205 raportów (1034 segmenty) opisujących takie zdarzenia. Następnie na ich podsawie zaprojektowano odpowiedni SI, dokonano ekstrakcji informacji oraz przeprowadzono analizę zebranych w ten sposób informacji.

5. Wnioski

Całość powyżej opisanego realizowanego procesu z przykładami składa się na przedstawiany w pracy oryginalny pomysł zintegrowanego rozwiązania problemu projektowania SI na podstawie analizy danych tekstowych. Rozwiązanie w postaci opracowanej przez autora metody przedstawiono na przykładzie analizy przypadku dotyczącego przetwarzania dokumentacji ze zdarzeń sporządzanej przez służby ratownicze PSP. Za swoje najważniejsze osiągnięcia autor uważa:

— opracowanie metody projektowania SI opartego o eksploracyjną analizę danych tekstowych,
— zrealizowanie przykładowego projektu informatycznego będącego studium przypadku realizacji autorskiej metody. Projekt ten był reakcją na pewne braki w dziedzinie ratownictwa. Braki te dotyczą ograniczonych i niewystarczających rozwiązań, w postaci metod i procesów analitycznych oraz aplikacyjnych, do badania zebranej w tym obszarze dokumentacji tekstowej,
— zaproponowanie jakościowej analizy tekstów w postaci zmodyfikowanej przez autora analizy przyczyn i skutków błędów z elementami analizy drzewa błędów oprogramowania do oceny tekstów pod kątem wyszukiwania z nich informacji i projektowania na jej podstawie modelu wybranego SI,
— włączenie powyżej wymienionej analizy do cyklu odkrywanie wiedzy z baz danych jako początkowego elementu służącego do zapoznania się z procesem akwizycji i zapisu informacji oraz celami eksploracji danych tekstowych,
— zaproponowane rozwiązania projektowe i implementację trzech systemów ekspertowych niezbędnych do zrealizowania toru formowania informacji i projektowania SI w postaci: SR, SKSS, SEIt,
— osiągnięcie dobrych wyników segmentacji tekstów na poziomie 95.5% F-miary lepszych od istniejących dotychczas rozwiązań dających wyniki na poziomie 87-88.7% F-miary,
— eksperymenty z doborem klasyfikatorów jak i ze skonstruowanym własnym rozwiązaniem, służących do procesu klasyfikacji krótkich form tekstowych w postaci segmentów,

— osiągnięcie zadowalających rezultatów klasyfikacji sięgających 90% F-miary dla klasyfikatora k-najbliższych sąsiadów,
— wykazanie, że obiegowa opinia o zastosowaniu n-gramów w procesie klasyfikacji polepsza jej parametry nie do końca się sprawdza w przypadku badanych tekstów. Zastosowanie bazy n-gramów w przypadku zastosowania klasyfikatora k-najbliższych sąsiadów pogarsza wyniki klasyfikacji o 5% a więc daje rezultaty w przybliżeniu równe 85% F-miary,
— pokazanie, że stosując prosty filtr lingwistyczny złożony tylko z lematyzatora można otrzymać zredukowany o połowę zbiór wyrażeń zachowując przy tym dobre wskaźniki klasyfikacji, które sięgają 90% F-miary dla klasyfikatora k-najbliższych sąsiadów,
— wprowadzenie filtru lingwistycznego powodującego redukcję cech o połowę podnosi znacznie jakość klasyfikacji za pomocą klasyfikatora Bayesa z 84.5% F-miary (niezredukowana przestrzeń cech) do 88% F-miary,
— eksperymenty nad doborem wag i miarami podobieństwa w reprezentacji wektorowej segmentów stosowanymi w procesie klasyfikacji. Eksperymenty te wykazały, że miary podobieństwa Kosinusowa, Jacarda oraz Dice dają znacznie lepsze wyniki klasyfikacji niż miara Euklidesowa. W skrajnych przypadkach wyniki są lepsze nawet o 20% F-miary. Zastosowanie natomiast schematu ważenia typu częstość termów - odwrotna częstość w dokumentach (*ang. term frequency-inverse document frequency - TF-IDF*) znacznie wpływa na proces klasyfikacji za pomocą klasyfikatora centroidalnego. Wyniki klasyfikacji polepszają się o około 4% F-miary w odniesieniu do innych schematów ważenia,
— zaprezentowanie możliwości ekstrakcji informacji z wybranej klasy segmentów do utworzonego SI za pomocą autorskiego rozwiązania. Wydobyto informacje i przytoczono statystyki na temat m.in. sprawności hydrantów czy też ich lokalizacji. W przypadku lokalizacji stwierdzono, że 31.14% opisów nie zawiera informacji o ich położeniu. Pozostałe 68.86% opisów informuje o tym, że hydrant znajdował się przy ulicy lub skrzyżowaniu,
— wykorzystanie zaproponowanej metody projektowania SI oraz opracowanego procesu ekstrakcji informacji do analizy wypadków samochodowych. W tym dodatkowym studium przypadku otrzymano statystyki oraz opisy: rodzaju marek samochodów które brały udział w wypadkach, sprzętu jaki najczęściej był używany w tego rodzaju interwencjach, jednostek PSP wyjeżdżających do danego rodzaju zdarzeń oraz rodzaju podejmowanych działań przez te jednostki,
— usystematyzowanie pojęć zebranych wokół eksploracyjnej analizy danych tekstowych,
— utworzenie podczas realizacji studium przypadku adnotowanego korpusu (segmenty przydzielone do klas semantycznych) z interwencji PSP wraz z jego opisem ilościowym w postaci odpowiednich wskaźników.

Wszystkie ww. elementy składają się na oryginalne, całościowe i zintegrowane rozwiązanie problemu projektowania systemu informacyjnego na podstawie analizy danych tekstowych. Tym samym omówione i zrealizowane aspekty badań udowadniają postawioną tezę i sprawiają, że *„Eksploracyjna analiza danych tekstowych może być zastosowana na etapie projektowania systemu informacyjnego stanowiącego narzędzie do realizacji procesów informacyjnych”*.

Literatura

[1] 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.

[2] Abakus: System EWID99. [on-line] [dostęp: 20 grudnia 2007]. Dostępny w Internecie: `http://www.ewid.pl/?set=rozw_ewid&gr=roz`.

[3] Abakus: System EWIDSTAT. [on-line] [dostęp: 20 grudnia 2007]. Dostępny w Internecie: `http://www.ewid.pl/?set=ewidstat&gr=prod`.

[4] Strona firmy Abakus. [on-line] [dostęp: 20 grudnia 2007]. Dostępny w Internecie: `http://www.ewid.pl/?set=main&gr=aba`.

[5] Marcin Mirończuk. Zmodyfikowana analiza FMEA z elementami SFTA w projektowaniu systemu wyszukiwania informacji na temat obiektów hydrotechnicznych w nierelacyjnym katalogowym rejestrze. *Studia Informatica*, 2(2B (97)):155–177, 2011.

[6] Solka Jeffrey. Text data mining: Theory and methods. *Statistics Surveys*, 2:94–112, 2008.

[7] Feldman Ronen, Sanger James. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2006.

[8] Michael W. Berry, Jacob Kogan. *Text Mining: Applications and Theory*. John Wiley & Sons, 2010.

[9] Kjersti Aas, Line Eikvil. Text categorisation: A survey, 1999.

[10] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

[11] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth. Advances in knowledge discovery and data mining. strony 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[12] William J. Frawley, Gregory Piatetsky-Shapiro, Christopher J. Matheus. Knowledge discovery in databases: An overview. strony 1–30, 1991.

[13] David Hand, Heikki Mannila, Padhraic Smyth. *Principles of Data Mining*. MIT Press, Cambridge, MA, 2001.

[14] Marcin Mirończuk. Przegląd i klasyfikacja zastosowań, metod oraz technik eksploracji danych. *Studia i Materiały Informatyki Stosowanej SIMIS*, 2(2):36–46, 2010.

[15] Ning Zhou, Hongli Cheng, Hongqin Chen, Shuang Xiao. The framework of text-driven business intelligence. *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on 21-25 Sept. 2007*, strony 5468–5471, 2007.

[16] Agnieszka Mykowiecka. *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym*. Warszawa: PJWSTK, 2007.

[17] Bogdan Stefanowicz. *Przestrzeń komunikatów. Przestrzeń informacyjna. Informacja*. Warszawa: Oficyna Wydawnicza Szkoła Główna Handlowa, 2010.

[18] Bogdan Stefanowicz. *Informacja*. Warszawa: Oficyna Wydawnicza Szkoła Główna Handlowa, 2010.

[19] Marcin Mirończuk, Tadeusz Maciak. Proces i metody eksploracji danych tekstowych do przetwarzania raportów z akcji ratowniczo-gaśniczych. *Metody Informatyki Stosowanej*, 4, 2011.

[20] Mariusz Flasiński. *Wstęp do sztucznej inteligencji*. Warszawa: PWN, 2011.

[21] Wiesław Traczyk. *Inżynieria Wiedzy*. Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2010.

[22] Moens Marie-Francine. *Information Extraction: Algorithms and Prospects in a Retrieval Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[23] Marcin Mirończuk. Wykorzystanie formalnej analizy pojęć do analizy dziedzinowych danych tekstowych. *Biuletyn WAT*, 2012.

[24] Wolff Karl Erich. A first course in formal concept analysis. *Context*, 93(3):429–438, 1993.

[25] Poelmans Jonas, Elzinga Paul, Viaene Stijn, Dedene Guido. Formal concept analysis in knowledge discovery: a survey. *Proceedings of the 18th international conference on Conceptual structures: from information to intelligence*, ICCS'10, strony 139–153, Berlin, Heidelberg, 2010. Springer-Verlag.

[26] Uta Priss. Formal concept analysis in information science. *Annual Review of Information Science and Technology (ARIST)*, 40:521–543, 1996.

[27] Adam Przepiórkowski. *Powierzchniowe przetwarzanie języka polskiego*. Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2008.

[28] Marcin Miłkowski, Jarosław Lipski. Using SRX standard for sentence segmentation. *LTC*, strony 172–182, 2009.

[29] opennlp. [on-line] [dostęp: 20 czerwca 2011]. Dostępny w Internecie: `http://incubator.apache.org/opennlp/`.

[30] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[31] Han Eui-Hong, Karypis George. Centroid-based document classification: Analysis and experimental results. *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, PKDD '00, strony 424–431, London, UK, UK, 2000. Springer-Verlag.

[32] Anna Kotulla. Wyszukiwanie informacji z uwzględnieniem danych dotyczących lokalizacji. *Studia Informatica*, 32(2B (97)):73–84, 2011.

Finansowanie

Praca naukowa współfinansowana ze środków Europejskiego Funduszu Społecznego, środków Budżetu Państwa oraz ze Środków Budżetu Województwa Podlaskiego w ramach projektu „Podlaska Strategia Innowacji – budowa systemu wdrażania".

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Indeks