STUDIA OECONOMICA POSNANIENSIA 2013, vol. 1, no. 5 (254)

Jacek Małyszko[1], Elżbieta Bukowska[1], Agata Filipowska[1], Bartosz Perkowski[1], Piotr Stolarski[1], Karol Wieloch
[1]Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Informatyki Ekonomicznej
Autor do korespondencji: Jacek Małyszko, jacek.malyszko@ue.poznan.pl

AUTOMATYCZNE ROZPOZNAWANIE OFERT KUPNA, SPRZEDAŻY I ZAMIANY W TEKSTACH W JĘZYKU POLSKIM[1]

Streszczenie: Artykuł prezentuje wyniki prac i eksperymentów dotyczących problemu przetwarzania niestrukturyzowanych tekstów napisanych w języku polskim w celu identyfikacji w nich ofert kupna, sprzedaży lub wymiany. W badaniach wykorzystano reguły ekstrakcji przygotowane na podstawie przeprowadzonej analizy korpusu. W artykule omówione są wybrane przykłady reprezentujące trudności, jakie niesie ze sobą omawiany problem. Opracowane podejście zostało poddane eksperymentalnej ocenie, na której podstawie skuteczność identyfikacji ofert została określona na 83% (według miary F1), natomiast określanie typu oferty (czy jest to kupno, czy sprzedaż) działa poprawnie w ponad 95% przypadków.

Słowa kluczowe: przetwarzanie języka naturalnego, ekstrakcja informacji.

Klasyfikacja JEL: C88, D83.

Wstęp

Pomimo realizacji wielu inicjatyw mających urzeczywistnić wizję internetu semantycznego, sieć ciągle jest źródłem informacji, w którym wiele zasobów pozostaje w niestrukturyzowanej formie. Liczne dostępne w internecie treści są formułowane wyłącznie w języku naturalnym i nie mają odpowiednich znaczników czy sformalizowanych opisów, dzięki którym możliwe byłoby ich automatyczne przetwarzanie (wizję takiej sieci, w której informacja miałaby jasno zdefiniowane znaczenie, co

[1] Praca naukowa finansowana ze środków na naukę w latach 2010–2012 jako projekt rozwojowy. Numer umowy: 0079/R/T00/2010 z dnia 29.10.2010 r.

ułatwiałoby jej automatyczne przetwarzanie, przedstawiono w artykule [Berners-Lee i in. 2001]). Obszarem zainteresowania artykułu jest zagadnienie związane z opisanym problemem, dotyczące przetwarzania niestrukturyzowanych tekstów w celu wykrycia, które z nich są ofertami kupna, sprzedaży bądź wymiany.

Przetwarzanie ofert kupna i sprzedaży znajduje szerokie pole zastosowań, przykładowo w wyszukiwarkach ogłoszeń czy też porównywarkach cenowych (usługach porównujących ceny tych samych towarów w różnych sklepach internetowych). Problem ten napotkany został również w projekcie Semantyczny monitoring cyberprzestrzeni (SMC)[2], mającym na celu opracowanie prototypu narzędzia, które umożliwiłoby monitorowanie źródeł internetowych w celu wykrycia potencjalnych zagrożeń w cyberprzestrzeni, manifestujących się w tych źródłach. Jako przypadek użycia analizowany w projekcie wybrano nielegalny handel lekami w internecie. Projekt SMC ma na celu monitorowanie określonych witryn internetowych, takich jak publicznie dostępne fora czy portale z ogłoszeniami, w celu wykrywania przypadków nielegalnego obrotu lekami w sieci (na przykład handlu środkami niedopuszczonymi do obrotu w kraju). Jednym z problemów badawczych napotkanych w trakcie prac nad projektem było właśnie rozpoznawanie ofert kupna, sprzedaży i wymiany na podstawie analizy tekstów napisanych w języku naturalnym.

W wielu sytuacjach, na przykład na portalach aukcyjnych czy na wyspecjalizowanych portalach z ogłoszeniami, oferty kupna i sprzedaży są w odpowiedni sposób oznaczone, dzięki czemu rozpoznawanie, czy dana wypowiedź jest ofertą, jest bezproblemowe. Przykładowo, często podczas tworzenia ogłoszenia jego autor zamieszcza je w odpowiednim dziale na stronie internetowej, co stanowi jednoznaczną wskazówkę dotyczącą typu treści zawartej w wypowiedzi. Jednocześnie oferty kupna czy sprzedaży zamieszczane są również w innych miejscach niż wyspecjalizowane portale, np. na forach internetowych, w komentarzach pod artykułami itp. W przypadku ogłoszeń dotyczących nielegalnego handlu lekami takie właśnie źródła są najpopularniejszym miejscem zamieszczania ofert. W ogłoszeniach tego typu cała informacja zawarta jest jedynie w wypowiedzi autora sformułowanej w języku naturalnym. Aby umożliwić automatyczne odnajdywanie określonych typów ofert w takich źródłach, najpierw konieczne jest określenie, czy dana wypowiedź zawiera ofertę, a następnie ustalenie, czy dotyczy ona kupna, sprzedaży, czy wymiany.

Artykuł ma na celu zaprezentowanie wyników analizy przeprowadzonej podczas prac nad rozpoznawaniem ofert kupna, sprzedaży lub wymiany dla potrzeb projektu SMC, a także omówienie przyjętego rozwiązania i zaprezentowanie poprawności wypracowanej metody. Warto podkreślić, że zaprezentowane rozwiązanie jest na bieżąco wykorzystywane w module ekstrakcji leksykalnej w systemie będącym wynikiem prac nad projektem Semantyczny Monitoring Cyberprzestrzeni. Artykuły ze zdefi-

[2] Projekt był realizowany przez Uniwersytet Ekonomiczny w Poznaniu (reprezentowany przez Katedrę Informatyki Ekonomicznej, której pracownikami są autorzy artykułu) oraz Sygnity S.A., *http://smc.kie.ue.poznan.pl/pl/* [dostęp 13 czerwca 2013].

niowanych źródeł internetowych (np. for internetowych czy portali z ogłoszeniami) są przez ten moduł przetwarzane, a jego wyniki wykorzystywane w celu przypisania stopnia zagrożenia do poszczególnych artykułów. Taki stopień zagrożenia jest wykorzystywany przez system do ustalenia, które artykuły powinny być prezentowane użytkownikom systemu (śledczym).

1. Analiza istniejących rozwiązań

Postawiony we wprowadzeniu problem można rozumieć jako problem klasyfikacji dokumentów tekstowych. Posiadając zdefiniowane klasy dokumentów (w naszym przypadku dokumenty zawierające oferty i niezawierające ofert), nowe dokumenty możemy przypisywać do jednej lub wielu klas [Sebastiani 2002]. Pomimo bogatej literatury z tego zakresu nie udało się odnaleźć ani polskojęzycznych, ani angielskojęzycznych publikacji prezentujących wyniki prac mających podobny cel do przyjętego w artykule, czyli określania, czy dokumenty są ofertami kupna, sprzedaży bądź zamiany. W niniejszej sekcji zaprezentowany zostanie krótki przegląd metod wykorzystywanych w klasyfikacji dokumentów tekstowych, które można użyć do rozwiązania postawionego problemu.

Jedną z technik stosowanych do klasyfikacji dokumentów jest uczenie maszynowe [Sebastiani 2002]. Przykładem metod tego typu są np. naiwne klasyfikatory bayerowskie [Frank i Bouckaert 2006] czy też maszyny wektorów wspierających [Joachims 1998]. Wspomniane podejścia działają na zasadzie analizy statystycznej polegającej na analizie częstości występowania poszczególnych słów w dokumentach z danej klasy, wyrażonej na przykład za pomocą miary *tfidf* [Sebastiani 2002]. Przykładowo, dla języka polskiego w pracy [Wawer 2011] za pomocą maszyn wektorów wspierających (*support vector machines*) wykrywano, jakiego atrybutu pewnego produktu dotyczy określony fragment tekstu. Innym przykładem klasyfikacji tekstu z wykorzystaniem uczenia maszynowego było określanie polarności opinii wyrażonej w danym zdaniu: czy opinia wyrażona w zdaniu jest pozytywna, czy negatywna [Wilson, Wiebe i Hoffmann 2009].

Inną techniką możliwą do wykorzystania w klasyfikacji tekstu może być użycie reguł, które mają za zadanie identyfikowanie (adnotowanie) fragmentów tekstu dotyczących danego tematu. Reguły takie operują na szablonach uwzględniających relacje pomiędzy słowami oraz semantycznymi klasami słów [Soderland 1999]. Reguły mogą być generowane na podstawie automatycznej analizy zaadnotowanego korpusu [Soderland 1999] lub też tworzone ręcznie [Pham i Pham 2012]. W ostatnich latach przetwarzanie dokumentów z wykorzystaniem reguł stosowano m.in. do ekstrakcji relacji przestrzennych [Zhang i in. 2009], ekstrakcji informacji z dokumentów klinicznych [Mykowiecka 2009], identyfikacji wymagań dla projektów informatycznych wyrażonych na forach internetowych [Vlas 2012] czy też ekstrakcji informacji z reklam nieruchomości (np. lokalizacji, ceny, typu nieruchomości itp.) [Pham i Pham 2012].

Istotą wspomnianych podejść jest przypisanie zidentyfikowanym fragmentom tekstu znaczenia zgodnie z zasadami zapisanymi w regule, które w określonych przypadkach można traktować jako przypisanie dokumentu do danej kategorii. Przykładami oprogramowania wykorzystywanego w tym celu są GATE[3] oraz SProUT[4] .

2. Analiza przykładów

W celu dokładniejszego zrozumienia istoty problemu poruszanego w artykule warto przeanalizować wybrane fragmenty ogłoszeń będących ofertami, prezentujące różnego rodzaju trudności, jakie napotkano podczas prac. Fragmenty te zaprezentowano w tabeli 1. Poniżej przeprowadzono analizę trudności, jakie przykłady te stawiają dla automatycznego przetwarzania.

Tabela 1. Przykłady rzeczywistych ofert kupna i sprzedaży

(1)	*Sprzedam* Faxigen XL 75mg
(2)	witam, czy ktoś *ma do sprzedania* MERIDIE 15 – tylko meridie nie zamiennik
(3)	Witam, *mam do sprzedania* 2 opakowania po 12szt. Malarone – GlaxoSmithKline
(4)	STREFA TANIEJ IZOTRETINOINY DOUSTNEJ – Izotek / Roaccutane / Aknenormin / Curacne. *mam* 2 opak gg 111111111
(5)	Tabletki sprawdzone, w 100% SKUTECZNE! Potrzebujesz pomocy, masz pytania, nie krępuj się pisz o każdej porze dnia czy nocy.
(6)	RE: *SPRZEDAM* SIBOTRIM 15. *jestem zainteresowana* prosze o kontakt przyklad@serwer.pl

Kursywą zaznaczono manualnie zaadnotowane fragmenty tekstu (dane kontaktowe zostały zanonimizowane)

Źródło: Na podstawie dokumentów pobranych z polskich for internetowych.

Część ofert jest bardzo łatwa do rozpoznania i ustalenia, czy dotyczy kupna, czy sprzedaży. Przykład (1) z tabeli 1 jest jednym z takich tekstów. W ogłoszeniu zawarto słowo oznaczające sprzedaż w pierwszej osobie liczby pojedynczej, co jednoznacznie wskazuje na to, że autor tekstu chce coś sprzedać. Takie oferty można łatwo rozpoznawać, porównując wszystkie wyrazy z przygotowaną listą słów oznaczających sprzedaż i kupno.

Trudniejszą sytuację napotykamy w przykładach (2) i (3). W obu przykładach pojawia się ten sam wyraz *sprzedania*, jednak w jednym przypadku mamy tu do czynienia z ofertą kupna (2), a w drugim z ofertą sprzedaży (3). Proste przeszukiwanie słownikowe pozwoli nam rozpoznać, że w tekstach tych zawarte są pewne oferty, jednak nie będziemy w stanie ustalić, czy dotyczą one kupna, czy sprzedaży. Konieczne

[3] http://gate.ac.uk/.

[4] http://sprout.dfki.de/.

jest tu przeanalizowanie kontekstu, w jakim wyraz *sprzedania* się znajduje. Różnicą jest tutaj to, że w przypadku (2) mamy do czynienia z pytaniem, za pomocą którego autor tekstu szuka innego podmiotu posiadającego w sprzedaży pewien obiekt (wskazuje na to słowo *ktoś* oraz słowo *ma*, będące odmienionym w trzeciej osobie liczby pojedynczej wariantem słowa *mieć*), stąd jesteśmy w stanie zrozumieć, iż jest to oferta kupna. Natomiast w przypadku (3) słowo *mam* implikuje fakt, że sformułowanie *do sprzedania* łączy się z pierwszą osobą liczby pojedynczej, czyli z autorem tekstu. Etap rozstrzygania z jakiego typu ofertą mamy do czynienia nazywamy normalizacją i zostanie on opisany w sekcji Opracowane rozwiązanie, w podsekcji Normalizacja.

Przykład (4) jest jeszcze bardziej skomplikowany. Brak tu jakiegokolwiek słowa jasno wskazującego na akt kupna lub sprzedaży, natomiast zaadnotowane zostało słowo *mam*. Dla człowieka oczywiste jest, że przykład ten jest ofertą sprzedaży, jednak proste wyszukiwanie słownikowe po słowie *mam* spowodowałoby zwrócenie bardzo dużej liczby tekstów niebędących ofertami (*mam* może być używane w wielu innych kontekstach, niż tylko oferta sprzedaży, na przykład w zdaniu *Mam grypę i muszę leżeć w łóżku*). Czynnikami, które wskazują tutaj na chęć sprzedaży, są nazwa obiektu sprzedaży i pojawiające się dane kontaktowe, które najprawdopodobniej służą do ustalenia szczegółów transakcji.

Jeszcze trudniejszym problemem jest rozpoznanie oferty w przykładzie (5). Została ona sformułowana w tak zawoalowany sposób, że brak w niej jednoznacznych przesłanek, iż mamy tu do czynienia z ofertą kupna bądź sprzedaży. Jedyną wskazówką dotyczącą tego, że omawiany wycinek tekstu jest ofertą sprzedaży, jest słowo *potrzebujesz*, jednak jednoznaczne rozstrzygnięcie tego, czy fraza ta powinna zostać uznana za ofertę, jest bardzo trudne (lub nawet niemożliwe) nawet dla człowieka. Fraza ta wskazuje, że możliwe jest sformułowanie ogłoszeń w sposób bardzo trudny do automatycznego rozpoznania.

Ostatni przykład (6) łatwo może być wykryty jako oferta, jednak poważną trudność stanowi poprawne rozpoznanie typu oferty (czy jest to sprzedaż, czy kupno). Tekst ten składa się z tytułu (cześć przed pierwszą kropką) i części właściwej. W tytule znajduje się wyraźne wskazanie na chęć sprzedaży, jednak samo ogłoszenie jest odpowiedzią wskazującą na chęć kupna. Wskazuje na to słowo *RE*: w tytule, które na forach internetowych używane jest w celu wskazania, że dana wypowiedź jest elementem pewnej dyskusji. W związku z tym ogłoszenie powinno zostać zaklasyfikowane jako oferta kupna.

3. Eksperyment i dostępne dane

Podejściem przyjętym w prowadzonych badaniach do rozwiązania opisywanego problemu jest wykorzystanie odpowiednio skonstruowanych leksykonów oraz reguł przetwarzania. To podejście do klasyfikacji wybrano ze względu na fakt, że rozpoznanie dokumentu jako oferty wymaga bardzo precyzyjnego przetworzenia wycinka tekstu, który zawiera słowa często wykorzystywane w ofertach. Podejścia wykorzystujące tyl-

ko częstości występowania słów w tekście, bez analizy kontekstu i zależności pomiędzy takimi słowami, w przypadku kilku przykładów omówionych w poprzedniej sekcji dałyby błędne rezultaty. Natomiast reguły ekstrakcji umożliwiają głębsze zrozumienie takich fragmentów tekstów, dzięki czemu możliwe byłoby rozstrzyganie przynajmniej części takich problemów. Dokumenty klasyfikowano właśnie na podstawie zidentyfikowanych (zaadnotowanych) fragmentów tekstu i rozpoznanych w nich znaczeń (przykładowo, jeśli w dokumencie rozpoznano fragment tekstu, w którym oferowana jest sprzedaż, cały dokument traktowano jako ofertę sprzedaży). Opracowywane rozwiązania poddawane były eksperymentalnej ocenie mającej na celu określenie poprawności rozpoznawania ofert kupna, sprzedaży i zamiany, dzięki czemu możliwe było opracowanie jak najlepszej metody w ramach obranego podejścia.

Podsumowując, automatyczne rozpoznawanie ofert polegało na adnotacji fragmentów tekstu, które wskazują na fakt, że dany dokument tekstowy jest ofertą kupna, sprzedaży lub zamiany (przykładowo w tabeli 1 takie fragmenty wyróżniono kursywą). Dodatkowo każda taka adnotacja zawierała informację o tym, który z trzech wspominanych typów ofert (kupna, sprzedaży lub zamiany) został rozpoznany. Eksperymenty polegały na porównywaniu takich adnotacji z adnotacjami przygotowanymi manualnie przez ekspertów i określenia zgodności adnotacji.

Eksperymenty, których rezultaty zaprezentowano w artykule, przeprowadzono na korpusie 452 dokumentów tekstowych pozyskanych dla potrzeb projektu SMC z wybranych źródeł, takich jak fora internetowe czy portale z ogłoszeniami. Każdy dokument był rzeczywistym wpisem lub ogłoszeniem na takiej witrynie, pobranym ze strony, na której pojawiła się co najmniej jedna oferta kupna lub sprzedaży, ale sam niekoniecznie jest taką ofertą. Na przykład w wątku na forum mogła pojawić się jedna oferta, a pozostałe wypowiedzi mogły dotyczyć innego tematu, a mimo to również zostały pobrane. Stąd korpus zawiera zarówno dokumenty zawierające oferty kupna, sprzedaży czy zamiany, jak i potencjalnie teksty na zupełnie inny temat. Wszystkie dokumenty zostały przeanalizowane przez ekspertów i napotkane słowa lub sformułowania wyrażające oferty zostały ręcznie zaadnotowane. W sumie zaadnotowano 444 takie sformułowania (w jednym dokumencie mogło być więcej niż jedno sformułowanie wskazujące na fakt, że jest to oferta).

Poprawność automatycznych adnotacji (czyli ich zgodność z adnotacjami manualnymi) była określana z wykorzystaniem standardowych miar stosowanych w ekstrakcji informacji, to jest precyzji, pełności i miary F1, statystyki obliczane były dla dwóch wariantów: ścisłego (*strict*) i łagodnego (*lenient*). W przypadku ścisłym automatyczna adnotacja traktowana jest jako poprawna tylko wówczas, gdy obejmuje dokładnie ten sam fragment tekstu co adnotacja ręczna. W przypadku łagodnym adnotacja jest poprawna zawsze, gdy obejmuje jakikolwiek wspólny zakres tekstu z adnotacją manualną. Rozróżnienie to jest istotne, ponieważ zaadnotowanie pełnego zakresu tekstu jest ważne dla poprawnego rozpoznawania typu oferty (czy jest to kupno, czy sprzedaż). Taki przypadek przedstawiono w tabeli 1 (przykłady (2) i (3)), gdzie bez analizy całości manualnej adnotacji nie da się prawidłowo określić typu oferty.

Dla potrzeb eksperymentów korpus podzielono na dwa zestawy: uczący i testowy. Zestaw uczący wykorzystany był do analizy sposobów formułowania ofert, natomiast zestaw testowy miał posłużyć do oceny skuteczności wypracowanej metody. Zestaw uczący zawierał 225, a testowy 219 adnotacji.

4. Opracowane rozwiązanie

W niniejszej sekcji szczegółowo opisane zostanie opracowane rozwiązanie. Podejściem przyjętym do rozwiązania opisywanego problemu w prowadzonych badaniach było wykorzystanie odpowiednio skonstruowanych leksykonów oraz reguł. Oprogramowaniem wykorzystywanym dla potrzeb eksperymentu było środowisko do przetwarzania języka naturalnego GATE oraz skrypty w języku Jython. Wybór został podyktowany faktem, że środowisko GATE jest wolnym oprogramowaniem, posiadającym bogatą dokumentację i z powodzeniem zastosowanym w szeregu projektów badawczych[5].

4.1. Leksykony

Pierwszym zasobem wykorzystywanym w przetwarzaniu są leksykony zawierające wyrazy, które mogą sygnalizować chęć kupna bądź sprzedaży. Ich wystąpienia w analizowanych dokumentach były wyszukiwane za pomocą mechanizmu ANNIE (*A Niearly-New Information Extraction System*) Gazetteer i odpowiednio adnotowane.

Na pierwszy, najprostszy leksykon składała się lista arbitralnie wybranych wyrazów, takich jak *kupić*, *mieć*, *sprzedać* wraz z ich możliwymi odmianami (pozyskanymi z generatora synonimów www.synomix.pl) oraz z wariantami będącymi błędnymi sposobami zapisu takich wyrazów (np. litera *ą* została zamieniona na om itd.). Tak wygenerowana lista posiadała ponad 5000 tysięcy wyrazów. Porównanie takiego leksykonu z korpusem uczącym ukazało znaczne jego niedopasowanie do naszych potrzeb. Wyszukiwanie za pomocą tego leksykonu wykazało, że w korpusie wystąpiło jedynie 81 wyrazów z niego pochodzących, a część z nich w rzeczywistych ogłoszeniach nie oznaczała oferty (np. wyrazy *sprzedająca*, *posiada* itp.). Do dalszych eksperymentów zredukowano więc listę do wyrazów, które zostały odnalezione w korpusie uczącym i przynajmniej w jednym przypadku rzeczywiście oznaczały ofertę (wyrazów takich było jedynie 48). W trakcie dalszej analizy poszerzono leksykon tak, że ostatecznie zawierał on 134 wyrazy. Celem takiego poszerzenia było zawarcie w nim wyrazów, które nie pojawiły się w analizowanym korpusie, jednak mają podobne znaczenie do wyrazów w leksykonie zawartych i potencjalnie również mogą być wykorzystywane w formułowaniu ofert.

[5] Listę takich projektów można znaleźć an stronie http://gate.ac.uk/projects.html [dostęp: 13 czerwca 2013].

4.2. Reguły

Kolejnym elementem prezentowanego rozwiązania są reguły analizujące kontekst, w jakim występują instancje słów z leksykonów. Dobrym przykładem uzasadniającym konieczność ich wykorzystania jest słowo *mam*, które pojawiło się w korpusie uczącym 80 razy, z czego 29 wystąpień rzeczywiście było powiązanych z pojawieniem się w tekście oferty (patrz przykład 4 w tabeli 1), a aż 51 wystąpień nie było z ofertą w żaden sposób powiązanych. Podobna sytuacja, choć w mniejszym zakresie, miała miejsce w przypadku słów: *sprzedać*, *wysyłam*, *kupić* czy *posiadam*. W celu odróżnienia wystąpień danych słów, które rzeczywiście oznaczają ofertę, od przypadków, gdy ich znaczenie jest odmienne, dokumenty przetwarzano z wykorzystaniem ręcznie przygotowanych reguł. Dodatkowo uwzględnienie kontekstu słów umożliwia przeprowadzanie głębszych analiz znaczeń fragmentów tekstu (co zostanie wykorzystane przy określaniu typu oferty, opisanym w kolejnej podsekcji).

W trakcie analizy ręcznie zaadnotowanego korpusu zidentyfikowano często występujące frazy wykorzystywane przy formułowaniu ofert:

– sformułowania typu *jestem zainteresowany/zainteresowana* połączone z niektórymi wyrazami z leksykonu słów wyrażających oferty, a dokładniej z rzeczownikami w narzędniku (np. *jestem zainteresowana kupnem*) lub też pytania o analogicznej konstrukcji (*Czy jest ktoś zainteresowany zakupem*...);
– sformułowaniami typu *mam do sprzedania* czy też *mam na sprzedaż*, gdzie pojawia się słowo *mam* i rzeczownik wyrażający ofertę w dopełniaczu lub bierniku; zamiast słowa *mam* mogą tu być też takie słowa jak *posiadam*, *chcę* czy też *mogę* w sformułowaniach *chcę kupić* (z czasownikiem w bezokoliczniku), *posiadam na sprzedaż* itp. Ponownie często spotykane są również pytania o analogicznej strukturze (np. *Jeśli ktoś miałby na sprzedaż, proszę o kontakt*...);
– sformułowania podobne do tych z poprzedniego punktu, lecz ze zmienionym szykiem, na przykład *do sprzedania mam*, *sprzedać mogę* oraz *do sprzedania.*

W każdym z tych przypadków, samo pojawienie się słowa z leksykonu może mieć różne znaczenie, jeśli jednak słowa te wykorzystane są zgodnie z ustalonymi regułami, z dużą pewnością możemy stwierdzić, iż rzeczywiście oznaczają one ofertę.

Dodatkowo zidentyfikowano inne sformułowania, które często mogą oznaczać ofertę pomimo tego, że nie ma w nich słów jednoznacznie wskazujących na sprzedaż, kupno lub wymianę. Dzieje się tak wówczas, gdy pojawiają się słowa lub frazy typu *mam*, *posiadam* czy też *jestem zainteresowana*, z występującymi niedaleko (np. w odległości do 5 wyrazów) nazwami obiektów często podlegających handlowi (w naszym przypadku leków), wyrażeniami wskazującymi na ilość lub danymi kontaktowymi (adresem e-mail, telefonem czy numerem komunikatora), jak np. w przykładzie (4) w tabeli 1.

Do każdego z omówionych powyżej typów sformułowań przygotowano odpowiednie reguły. Konieczne było również przygotowanie dodatkowych słowników oraz dokonanie podziału już istniejących leksykonów. W efekcie powstały dwie klasy słowników:

– podzielono wyrazy na osobne leksykony w zależności od ich części mowy, czasu dla czasowników, przypadku dla rzeczowników itd.; tak utworzoną klasę słowników będziemy dalej nazywać leksykonami akcji;
– wyodrębniono słowo *mam* do osobnego leksykonu, do którego dodano jeszcze takie wyrazy, jak *chcę*, *mogę* i *posiadam* wraz z odpowiednimi odmianami, przygotowano również leksykony słów wskazujących na to, że dane sformułowanie jest pytaniem (np. *czy*, *jeśli* itp.), co utworzyło klasę słowników zwaną dalej leksykonami uzupełniającymi.

Na tej podstawie przygotowano reguły służące do wyszukiwania w tekście sekwencji słów z odpowiednich słowników i w odpowiedniej formie. Przygotowano je z wykorzystaniem formalizmu JAPE (tzw. gramatyki JAPE – *Java Annotation Patterns Engine*, Silnik Wzorców Adnotacji dla języka Java). Reguły można podzielić na pięć typów:

1. Reguły wykrywające sformułowania zawierające wyrazy z leksykonów akcji będące w odpowiedniej formie i połączone w określonej kolejności z wyrazami z leksykonów uzupełniających, zwane regułami pełnymi; dodatkowo, za pomocą takich reguł ustalane jest również, czy dane sformułowanie nie jest pytaniem (co będzie przydatne w dalszej części przetwarzania w trakcie normalizacji).
2. Reguły zaznaczające następujące fragmenty tekstu: nazwy leków (na podstawie leksykonu takich nazw), ilości (np. *dwa opakowania*), adresy e-mail, numery telefonów oraz typy komunikatorów i odpowiednie numery lub pseudonimy użytkowników w ramach danego komunikatora.
3. Reguły adnotujące słowa z leksykonów uzupełniających wówczas, gdy w ich sąsiedztwie nie ma wyrazów z leksykonu akcji, ale występują tam inne typy adnotacji wspomniane w poprzednim punkcie, np. dane kontaktowe itp. (ten typ reguł dalej będzie nazywany regułami uzupełniającymi i ma za zadanie wykrywać oferty sformułowane w sposób podobny do przykładu (4) z tabeli 1).
4. Reguły adnotujące wszystkie pozostałe wystąpienia słów z leksykonów akcji niezależnie od ich kontekstu, dalej zwane regułami prostymi.
5. Reguła wykrywająca negacje przed adnotacjami utworzonymi z wykorzystaniem reguł opisanych powyżej i zatwierdzająca jako faktyczne oferty jedynie adnotacje niezanegowane.

4.3. Normalizacja

Ostatnim krokiem w rozpoznawaniu oferty jest odróżnianie ofert kupna od sprzedaży (etap ten zwany jest w naszym podejściu normalizacją)[6]. W wielu przypadkach jest to proste zadanie, jednak zdarzają się takie sformułowania, w których staje się ono kłopotliwe. Przykłady trudniejszych sformułowań przedstawiono w tabeli 1 w wierszach (2) i (3). W celu rozróżnienia takich przypadków dokonano następujących czynności:

[6] Należy zwrócić uwagę na fakt, że z uwagi na skonstruowane słowniki, rozpoznawanie ofert zamiany było jednoznaczne i te typy ofert nie były poddawane normalizacji.

– Leksykony akcji podzielono na osobne słowniki dla sprzedaży, kupna i zamiany, dzięki czemu pozyskano informację o tym, co zazwyczaj dane słowa oznaczają. Podziału dokonano poprzez analizę ręcznie zaadnotowanych przykładów. W niektórych przypadkach podział ten nie zawsze był oczywisty; przykładowo, wyrazy takie jak *kup*, *kupując* czy *zakup*, które na pierwszy rzut oka można by przyporządkować do leksykonu wyrazów oznaczających kupno, faktycznie częściej pojawiają się w tekstach zawierających oferty sprzedaży.
– Z leksykonów uzupełniających wyodrębniono wyrazy w odpowiedniej odmianie, których użycie zmienia znaczenie sformułowania na odwrotne. Wyrazy takie to m. in. *ma*, *miałby, masz* i *posiada*. Reguły adnotowały frazy z takimi wyrazami tylko wówczas, gdy wcześniej pojawiły się słowa wskazujące na to, że dana fraza jest pytaniem lub zdaniem warunkowym (czyli słowa takie, jak: *czy*, *jeśli*, *gdyby* itp.). Przykładem frazy, gdzie słowo *ma* zmienia znaczenie interpretacji całości ze sprzedaży na kupno, jest wiersz (2) z tabeli 1.

Normalizacja przebiegała w następujących krokach: najpierw przypisywano do adnotacji znaczenie na podstawie tego, z jakiego słownika pochodził wyraz z leksykonu akcji. Przykładowo w zdaniu *Jeśli ktoś chciałby odkupić to zapraszam*, słowo *odkupić* pochodzi z leksykonu słów domyślnie oznaczających kupno, takie też pierwsze znaczenie zostaje przypisane do całego sformułowania. Następnie, w przypadku reguł biorących pod uwagę słowa z leksykonów uzupełniających, znaczenie było odwracane w przypadku, gdy zaznaczany fragment tekstu został wykryty jako pytanie lub zdanie warunkowe. W podanym powyżej przykładzie rozpoznano zdanie warunkowe ze względu na wystąpienia tam słów *jeśli* oraz *chciałby*.

5. Wyniki i ich interpretacja

Wyniki uzyskiwane w trakcie prac nad rozwiązaniem zaprezentowano w tabeli 2. Jako rozwiązanie podstawowe – porównawcze (*baseline*) przyjęto wyszukiwanie słownikowe za pomocą leksykonu powstałego po analizie korpusu uczącego (patrz pierwszy wiersz w tabeli 2), bez zastosowania reguł. Wyszukiwanie za pomocą pierwszego, najprostszego leksykonu, na który składała się lista arbitralnie wybranych wyrazów używanych przy wyrażaniu ofert, pozwoliło uzyskać wyniki jedynie na poziomie nieznacznie przekraczającym 50% według miary F1, stąd pominięto je w tabeli.

Użycie gramatyk JAPE pozwoliło na znaczny wzrost poprawności rozpoznawania ofert, dochodzący do prawie 10 punktów procentowych względem punktu wyjścia. Warianty drugi i trzeci w tabeli 2 reprezentują statystyki poprawności dla różnych zestawów reguł. Wariant z wiersza drugiego zawiera tylko reguły wymagające wystąpienia słów z leksykonów akcji (czyli reguły typu 1, 4 i 5 z listy na poprzedniej stronie), natomiast zestaw z ostatniego wiersza zawiera również reguły uzupełniające (typ 3). Jak widać w tabeli 2, poszerzenie rozwiązania o reguły typu trzeciego pogarszało

nieznacznie precyzję, dając jednocześnie wzrost w pełności, przy czym w korpusie uczącym miara F1 wzrosła o prawie 2 punkty procentowe, natomiast w korpusie testowym nieznacznie zmalała (o około 1 punkt procentowy).

Tabela 2. Statystyki poprawności rozpoznawania ofert dla różnych testowanych wariantów

Wariant	Podstawa porównywania	Zestaw uczący		
		precyzja	pełność	F1
baseline	ścisły	0,6926	0,7417	0,7163
	łagodny	0,8521	0,9125	0,8813
reg. 1, 4 i 5	ścisły	0,8347	0,8625	0,8484
	łagodny	0,871	0,9	0,8852
reg. 1–5	ścisły	0,834	0,9	0,8657
	łagodny	0,8725	0,9416	0,9058
Wariant	**Podstawa porównywania**	**Zestaw testowy**		
		precyzja	**pełność**	**F1**
baseline	ścisły	0,6708	0,7488	0,7077
	łagodny	0,7792	0,8698	0,822
reg. 1, 4 i 5	ścisły	0,7489	0,8186	0,7822
	łagodny	0,8	0,8744	0,8355
reg. 1–5	ścisły	0,7236	0,8279	0,7739
	łagodny	0,7724	0,8837	0,8243

Na koniec warto przeanalizować statystyki poprawności adnotacji dla poszczególnych rodzajów reguł. Tabela 3 zawiera informacje o liczbie adnotacji, które zostały oznaczone za pomocą poszczególnych typów reguł (reguł pełnych, uzupełniających i prostych), i jaką osiągnięto dla nich precyzję. Dane wskazują, że reguły pełne odpowiadają za niecałe 20% wszystkich adnotacji, jednak działają z bardzo wysoką precyzją. W systemach wykorzystujących takie adnotacje możliwe jest traktowanie ich z większą ufnością. Natomiast dzięki regułom uzupełniającym możliwe jest rozpoznanie ofert niezawierających słów z leksykonów akcji z precyzją zbliżoną do reguł prostych.

Tabela 3. Ocena skuteczności poszczególnych typów reguł

Typ reguł	Liczba adnotacji	Precyzja	
		ścisła	łagodna
Pełne	39	0,9231	1
Uzupełniające	10	0,8	0,9
Proste	209	0,8278	0,8517

Normalizację oceniono wyłącznie dla tych sformułowań, które zostały poprawnie zaadnotowane jako oferty (błędnie rozpoznane słowa wyrażające ofertę nie mogą być dobrze znormalizowane). Takich fraz było 403. Normalizacja dała poprawne wyniki w 395 z nich, w związku z czym jej skuteczność można ocenić na 98%. Poprawność normalizacji dla poszczególnych typów ofert zaprezentowano w tabeli 4.

Tabela 4. Ocena skuteczności normalizacji dla poszczególnych typów ofert. Wartości na przekątnej tabeli odpowiadają poprawnym wynikom

		Wyniki normalizacji		
		Sprzedaż	**kupno**	**zamiana**
Adnotacje ręczne	**sprzedaż**	295	2	0
	kupno	6	97	0
	zamiana	0	0	3

Podsumowanie i dalsze prace badawcze

W artykule zaprezentowano regułowe podejście do automatycznego rozpoznawania ofert kupna, sprzedaży i zamiany. Wypracowane metody pozwoliły na uzyskanie miary F1 na poziomie przekraczającym 80% dla zestawu testowego, natomiast normalizacja (określanie, czy dana oferta dotyczy kupna czy sprzedaży) osiągnęła skuteczność na poziomie ponad 95%. Należy dodatkowo zwrócić uwagę na fakt, że na statystyki ma wpływ charakterystyka korpusu, na którym przeprowadzano badania, a dokładniej fakt, iż wiele tekstów zawierało literówki, błędy ortograficzne itp., przez co w niektórych przypadkach wypracowane metody nie były w stanie poprawnie rozpoznać ofert.

Omówione badania mogą być kontynuowane poprzez dodawanie kolejnych reguł. Konieczne jest m.in. określenie reguł sprawdzania, czy treść dokumentu nie jest odpowiedzią na inne ogłoszenie (zgodnie z przykładem (6) w tabeli 1) w celu poprawnego normalizowania takich przypadków, co nie było przedmiotem badań w niniejszym artykule. Ciekawym dalszym kierunkiem badań w omawianym zakresie byłoby rozpoznawanie ofert z wykorzystaniem uczenia maszynowego (przykładowo z wykorzystaniem zaadnotowanych przykładów w poszukiwaniu reguł asocjacyjnych), a nie manualnie przygotowywanych reguł.

Bibliografia

Berners-Lee, T., Hendler, J., Lassila, O. i in., 2001, *The Semantic Web. Scientific American*, 284(5), s. 28–37.

Frank, E. Bouckaert, R. (2006), *Naive Bayes for Text Classification with Unbalanced Classes*, Knowledge Discovery in Databases: PKDD 2006, s. 503–510.
Joachims, T., 1998, *Text Categorization with Support Vector Machines*: *Learning with Many Relevant Features*, Machine learning: ECML-98, s. 137–142.
Mykowiecka, A., Marciniak, M., Kupść, A., 2009, *Rule-based Information Extraction from Patients' Clinical Data*, Journal of biomedical informatics, vol. 42(5), s. 923–936.
Pham, L.V., Pham, S.B., 2012, *Information Extraction for Vietnamese Real Estate Advertisements*, Fourth International Conference on Knowledge and Systems Engineering (KSE), s. 181–186.
Sebastiani, F., 2002, *Machine Learning in Automated Text Categorization*, ACM Comput. Surv., vol. 34(1), s. 1–47.
Soderland, S., 1999, *Learning Information Extraction Rules from Semi-structured and Free Text*, Machine Learning, vol. 34(1–3), s. 233–272.
Vlas, R.E., Robinson, W.N, 2012, *Two Rule-based Natural Language Strategies for Requirements Discovery and Classification in Open Source Software Development Projects*, Journal of Management Information Systems, vol. 28(4), s. 11–38.
Wawer, A. (2011), *Mining Opinion Attributes From Texts Using Multiple Kernel Learning*, IEEE 11th International Conference on Data Mining Workshops.
Wilson, T., Wiebe, J, Hoffmann, P., 2009, *Recognizing Contextual Polarity*: *An Exploration of Features for Phrase-level Sentiment Analysis*, Computational linguistics, vol. 35(3), s. 399–433.
Zhang, C., Zhang, X, Jiang, W., Shen, Q., Zhang, S., 2009, *Rule-based Extraction of Spatial Relations in Natural Language Text*, International Conference on Computational Intelligence and Software Engineering, s. 1–4

AUTOMATIC IDENTIFICATION OF BUY, SELL AND EXCHANGE OFFERS IN UNSTRUCTURED TEXTS WRITTEN IN THE POLISH LANGUAGE

Abstract: This article presents the results of research and experimentation on processing unstructured texts written in the Polish language in order to identify which of these texts contain buy, sell or exchange offers. The approach applied was based on manually prepared rules of extraction based on an analysis of a corpus of documents obtained from the Internet (within the Semantic Monitoring of Cyberspace project). In the article, selected examples of text fragments are discussed which show what challenges had to be addressed to solve the problem. The chosen approach was then experimentally evaluated; the accuracy in identifying offers reaching 83% (according to the F_1-score), while determining the offer type (whether buying or selling) was correct in over 95% of cases.

Keywords: industrial organization, industry studies: services, information and internet services; computer software.