Zeszyty Naukowe
Wyższej Szkoły Bankowej w Poznaniu
Nr 40/2012

Dariusz Ceglarek
Wyższa Szkoła Bankowa w Poznaniu

Zastosowanie kompresji semantycznej w zadaniach przetwarzania języka naturalnego

Streszczenie. *Kompresja semantyczna jest techniką pozwalającą uzyskać właściwą generalizację pojęć w zależności od kontekstu, dzięki czemu można znaleźć w różnych dokumentach tę samą myśl inaczej sformułowaną lub sformułowaną z użyciem innych pojęć. Rozwój koncepcji kompresji semantycznej i opracowanie nowych algorytmów pozwolił zastosować ją do klasyfikacji dokumentów i rozbudowy struktur reprezentacji wiedzy, takich jak sieci semantyczne. W artykule przedstawiono wyniki badań nad nowymi metodami i narzędziami kompresji semantycznej, które zostały przystosowane do zadań przetwarzania języka naturalnego.*

Słowa kluczowe: *kompresja semantyczna, ochrona własności intelektualnej, przetwarzanie języka naturalnego, reprezentacja wiedzy, sieć semantyczna*

1. Wprowadzenie

Kompresja semantyczna została opracowana z myślą o sytuacjach, gdy dwa lub więcej dokumentów zawiera wspólne fragmenty z pewnymi modyfikacjami przeprowadzonymi z użyciem słowników czy tezaurusów (np. poprzez użycie pojęć synonimicznych), ale które nie są podobne do siebie w sensie dosłownego porównania słowo po słowie, tak jak postępuje się w systemach wyszukiwawczych (*information retrieval systems* – IR). Z sytuacją taką mamy do czynienia w zadaniu wykrywania plagiaryzmów w określonych korpusach dokumentów[1]. Badania nad kompresją

[1] Zagadnienie to zostało opisane w: T. Ota, S. Masuyama, *Automatic plagiarism detection among term papers*, w: *Proceedings of the 3rd International Universal Communication '09*, ACM, 2009, s. 395-399; R. Lukashenko, V. Graudina, J. Grundspenkis, *Computer-based plagiarism detection methods and tools: an overview*, w: *Proceedings of the 2007 International Conference on Computer Systems and Technologies, CompSysTech '07. New York, USA*, ACM, 2007, s. 401-406;

semantyczną zostały zainicjowane podczas prac autora nad systemem ochrony własności intelektualnej SOWI[2].

Zadaniem systemu SOWI jest ochrona własności intelektualnej zawartej w dokumentach tekstowych, w tym wykrywanie zapożyczeń – sprawdzenie, czy w danym dokumencie tekstowym występuje odpowiednio duży fragment tekstu, który pokrywa się z treścią innego dokumentu w takim stopniu, że można mówić o zapożyczeniu treści i naruszeniu własności intelektualnej. Wdrożono tu autorskie algorytmy, co sprawia, że metody sprawdzania fraz wspólnych w dokumentach są niewrażliwe na zabiegi osób chcących ukryć fakt zapożyczenia fragmentów tekstu poprzez zmiany szyku tekstu oraz stosowanie w dokumencie synonimów czy pojęć bliskoznacznych. Zastosowanie w tych metodach kompresji semantycznej spowodowało, że możliwe jest wykrywanie nie tylko zapożyczeń w formie dosłownego skopiowania fragmentu tekstu, ale także polegających na przedstawieniu tej samej myśli za pomocą innych sformułowań. Stało się to możliwe również dzięki wykorzystaniu zaawansowanych struktur reprezentacji wiedzy o języku naturalnym, jakimi są sieci semantyczne. Autor dostrzegł możliwość zastosowania idei kompresji semantycznej również w innych sytuacjach, poprzez odpowiednie dostosowanie narzędzi i rozwiązań, zgodnie ze specyfiką rozmaitych zadań w ramach przetwarzania języka naturalnego.

Kompresja semantyczna może być także cennym narzędziem w zadaniach, w których głównym celem przetwarzania informacji jest przedstawienie użytkownikowi informacji dopasowanej do jego indywidualnych wymagań. Kompresję semantyczną można zatem zdefiniować jako skuteczną technikę uogólniania pojęć, która dopasowuje się do kontekstu i uwzględnia dodatkowo wymóg minimalizowania straty informacyjnej.

Powyższa definicja podkreśla potrzebę określenia właściwego kontekstu dla każdego pojęcia, które pojawia się w przetwarzanym dokumencie. Jest to zadanie trudne i jedynie osoba dysponująca odpowiednią wiedzą jest w stanie podać ze stuprocentową skutecznością właściwe znaczenie każdego pojęcia, gdyż w procesie tym należy uwzględnić również konotacje kulturowe danego pojęcia.

Autor wykazał, że kompresja semantyczna daje dobre wyniki, prawidłowo określając formy generalizujące pojęcia w języku naturalnym. Po raz pierwszy idea kompresji semantycznej pojawiła się w pracy N.N. Percovej[3], autorka nie podała jednak sposobu jej realizacji.

S. Burrows, S.M.M. Tahaghoghi, J. Zobel, *Efficient plagiarism detection for large code repositories*, „Software: Practice and Experience" 2007, t. 37, nr 2, s. 151-175.

[2] D. Ceglarek, *Koncepcja komponentowego systemu ochrony własności intelektualnej wykorzystującego semantyczne struktury informacji*, w: *Technologie informatyczne w zarządzaniu wiedzą – uwarunkowania i realizacja*, red. P. Adamczewski, M. Zakrzewicz, Wyd. WSB w Poznaniu, Poznań 2009.

[3] N.N. Percova, *On the types of semantic compression of text*, w: *COLING '82. Proceedings of the 9th conference on Computational linguistics*, t. 2, Academia Praha, 1982, s. 229-231.

Pierwotny pomysł kompresji semantycznej, która umożliwiałaby prawidłowe ujednoznacznianie pojęć wieloznacznych[4] podczas procesu ich generalizowania, został opracowany przez autora tego artykułu w postaci algorytmu kompresji semantycznej[5]. Algorytm został następnie zaimplementowany oraz przetestowany w szeregu eksperymentów, których efektem było wiele ulepszeń i rozszerzeń w stosunku do pierwowzoru. Istotne właściwości uzyskanego algorytmu to:

– zdefiniowanie i zaprezentowanie kompresji semantycznej jako technologii przydatnej w zadaniach przetwarzania języka naturalnego; skonstruowanie i zaimplementowanie w algorytmie słowników frekwencyjnych, które w przypadku występowania pojęć wieloznacznych wraz z algorytmami określającymi właściwe hiperonimy pojęć w zależności od kontekstu informacyjnego[6],

– przekształcenie sieci semantycznej WordNet[7] do postaci WiSENet, co spowodowało, że eksperymenty określające jakość kompresji semantycznej stały się możliwe zarówno dla dokumentów w języku polskim, jak i w angielskim[8],

– wysoce specjalizowany automat skończony pozwalający automatyzować budowę reguł, które wydobywają nowe pojęcia oraz nowe relacje leksykalne[9].

Niniejszy artykuł stanowi podsumowanie przeprowadzonych już badań, zatem jego struktura powinna konsekwentnie porządkować ich rezultaty. Dlatego następna sekcja poświęcona jest reprezentacji wiedzy, zwłaszcza strukturom tej

[4] M. Sanderson, *Word Sense Disambiguation and Information Retrieval*, w: *SIGIR '94. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, red. W.B. Croft, C.J. van Rijsbergen, SIGIR, ACM/Springer, New York 1994, s. 142-151; J. Boyd-Graber, D.M. Blei, X. Zhu, *A Topic Model for Word Sense Disambiguation*, w: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, June 2007*, s. 1024-1033.

[5] D. Ceglarek, K. Haniewicz, W. Rutkowski, *Quality of Semantic Compression in Classification*, w: *Computational Collective Intelligence, Second International Conference, ICCCI 2010, Kaohsiung, Taiwan, November 10-12, 2010. Proceedings*, cz. 1, red. J.-S. Pan, S.-M. Chen, N.T. Nguyen, Springer-Verlag, Berlin – Heidelberg 2010, „Lecture Notes in Computer Science" 2010, t. 6421, s. 162-171.

[6] R. Snow, D. Jurafsky, A.Y. Ng, *Learning syntactic patterns for automatic hypernym discovery*, w: *Advances in Neural Information Processing Systems (NIPS)*, 2005. Dokładny opis tego mechanizmu można znaleźć w: D. Ceglarek, K. Haniewicz, W. Rutkowski, *Semantic Compression for Specialised Information Retrieval Systems*, w: *Advances in Intelligent Information and Database Systems*, red. N.T. Nguyen, R. Katarzyniak, S.-M. Chen, Springer Verlag, Berlin – Heidelberg 2010, „Studies in Computational Intelligence" 2010, t. 283, s. 111-121.

[7] M. Miłkowski, *Automated Building of Error Corpora of Polish*, w: *Corpus Linguistics, Computer Tools, and Applications – State of the Art*, PALC 2007, red. B. Lewandowska-Tomaszczyk, Peter Lang, Frankfurt am Main 2008, s. 631-639.

[8] Przekształcenie to zostało opisane w: D. Ceglarek, K. Haniewicz, W. Rutkowski, *Quality of Semantic Compression…*

[9] D. Ceglarek, K. Haniewicz, W. Rutkowski, *Towards Knowledge Acquisition with WiSENet*, w: *New Challenges for Intelligent Information and Database Systems*, red. N.T. Nguyen, B. Trawinski, J.J. Jung, Springer Verlag, Berlin – Heidelberg 2011, „Studies in Computational Intelligence" 2011, t. 351, s. 75-84.

reprezentacji, ze szczególnym naciskiem na sieci semantyczne. Następnie przedstawiono proces przekształcenia najbardziej popularnej sieci semantycznej dla języka angielskiego WordNet do formatu SenecaNet, czego efektem jest sieć semantyczna WiSENet. Kolejna sekcja poświęcona jest globalnej i dziedzinowej kompresji semantycznej. Przedstawiono tu algorytmy i mechanizmy służące do jej utworzenia oraz przykłady zastosowań – m.in. pokazano, że kompresja semantyczna użyta w zadaniu klasyfikacji dokumentów tekstowych metodami analizy skupień podnosi jakość klasyfikacji dokumentów. Następny przykład pokazuje, jak kompresja semantyczna może być użyta do rozbudowy samej sieci semantycznej, przez co rozumie się odkrywanie nowych pojęć w celu ich dodania do sieci oraz odkrywanie nowych relacji leksykalnych pomiędzy konceptami już zgromadzonymi w sieci semantycznej. Ostatni przykład dotyczy zastosowania kompresji semantycznej do wspomagania rozumienia tekstu poprzez dopasowanie prezentowanych użytkownikowi pojęć zgodnie z jego poziomem rozumienia tekstów z danej dziedziny. Artykuł kończy się podsumowaniem, konkluzjami oraz wskazuje kierunki przyszłych badań.

2. Reprezentacja wiedzy

Analizą i automatycznym wyodrębnianiem prawidłowości w zbiorach dokumentów tekstowych i tekstowych bazach danych zajmuje się *text mining*, który jest multidyscyplinarną dziedziną, wykorzystującą m.in. metody statystyczne, metody systemów wyszukiwawczych (*information retrieval*) czy maszynowe uczenie. Ogólniejszą dyscypliną obejmującą problematykę sztucznej inteligencji i językoznawstwa, zajmującą się automatyzacją analizy, rozumienia, tłumaczenia i generowania języka naturalnego, jest przetwarzanie języka naturalnego (*Natural Language Processing* – NLP).

Metody *text miningu* składają się zazwyczaj z dwóch etapów: wygładzania tekstu (*text refining*) oraz wydobywania wiedzy (*knowledge discovery*). Na etapie wygładzania tekstu pozbawiony struktury dokument tekstowy przekształcany jest w formę pośrednią[10], tworzoną w celu wykrycia zależności między dokumentami (wydobycie wiedzy) w drugim etapie z wykorzystaniem metod charakterystycznych dla danego zadania *text miningu*[11].

W przypadku wszystkich zadań realizowanych w ramach zadań przetwarzania języka naturalnego niezbędne jest przeprowadzenie wygładzania tekstu, które pole-

[10] Forma pośrednia może mieć postać sekwencji cech, wektora cech lub grafu konceptualnego.

[11] Typowe zadania w ramach *text miningu* obejmują klasyfikację dokumentów (grupowanie, kategoryzację), automatyczne streszczanie dokumentów, grupowanie pojęć, wizualizację i nawigację w zbiorze dokumentów i ekstrakcję informacji.

ga na przekształceniu wyjściowego dokumentu tekstowego w strukturę zawierającą ułożone sekwencyjnie deskryptory pojęć (konceptów) występujących w wyjściowym dokumencie. Na wygładzanie tekstu składają się operacje: wyodrębnienia wyrazów (*tokenization*), usunięcia słów niemających znaczenia informacyjnego z tzw. stop-listy, identyfikacji pojęć wielowyrazowych oraz lematyzacji pojęć.

Ostatnią fazą wygładzania tekstu jest ujednoznacznienie pojęć (*disambiguation*), czyli wyznaczenie właściwych pojęć dla wieloznacznych termów[12], które wystąpiły w dokumencie (eliminowanie tzw. pozornego podobieństwa). Zjawisko wieloznaczności pojęć, zwane polisemią, dotyczy każdego języka naturalnego i oznacza, że jednemu konceptowi odpowiada wiele znaczeń, czyli że różne pojęcia nazywane są tak samo. Przykładem polisemii może być koncept „dysk", który ma takie znaczenia, jak: „komputerowy nośnik pamięci", „przyrząd lekkoatletyczny", „kość składowa kręgosłupa" lub „owalny kształt". Celem zastosowania ujednoznaczniania pojęć jest lepsze odwzorowanie termów występujących w dokumentach we właściwe pojęcia, a zatem lepsze dopasowanie informacji wywnioskowanej z dokumentów do potrzeb informacyjnych[13]. Istnieje wiele metod ujednoznaczniania pojęć, w tym metody oparte na semantycznej reprezentacji wiedzy. Skuteczną[14] metodę ujednoznaczniania pojęć (o skuteczności na poziomie dochodzącym do 82%), wykorzystującą sieć semantyczną dla języka polskiego, zaproponował autor niniejszego opracowania[15]. Wykorzystuje ona zależności semantyczne między pojęciami w sieci semantycznej SenecaNet dla języka polskiego, a jej działanie opiera się na wskazaniu najbardziej prawdopodobnego znaczenia pojęcia wieloznacznego – biorąc pod uwagę lokalny kontekst użycia owego pojęcia w badanym dokumencie.

2.1. Struktury reprezentacji wiedzy

Metody reprezentacji wiedzy są sposobem, w jakim wiedza o świecie jest przedstawiana wraz z metodami jej przetwarzania i wnioskowania (inferencji). Jest to ściśle określony język opisu wiedzy zaopatrzony w mechanizm jej przetwarzania.

---

[12] Termem jest słowo lub wielowyrazowy związek semantyczny (związek frazeologiczny, kolokacja), który wystąpił w dokumencie.

[13] Ch. Stokoe, M.P. Oakes, J. Tait, *Word Sense Disambiguation in Information Retrieval Revisited*, SIGIR, 2003.

[14] Skuteczne metody prawidłowo identyfikują od 70 do 75% znaczeń pojęć wieloznacznych. W pracy M. Sanderson, *Retrieving with Good Sense*, „Informational Retrieval" 2000, t. 2, nr 1, s. 49-69, pokazano, że wyłącznie metody analizy lingwistycznej są w stanie pokonać poziom 90% skuteczności ujednoznaczniania pojęć wieloznacznych.

[15] D. Ceglarek, *Zastosowanie sieci semantycznej do disambiguacji pojęć w języku naturalnym*, w: *Systemy wspomagania organizacji SWO 2006*, Wyd. AE w Katowicach, Katowice 2006.

Każdemu pojęciu odpowiada w języku naturalnym zapis w postaci wyrazu, kolokacji[16] lub związku frazeologicznego, który jest jego odzwierciedleniem. Zapis pojęcia w języku naturalnym nazywamy konceptem. Celem jest stworzenie przetwarzalnej przez system reprezentacji dokumentu, tak aby na podstawie treści dokumentu wyodrębnić jednostki odpowiadające znaczeniu informacyjnemu konceptów.

Popularne w systemach wyszukiwawczych i *text miningu* metody opierają się na prostej strukturze reprezentacji wiedzy, gdzie dokumenty reprezentowane są przez zbiory słów kluczowych, a najbardziej popularnym modelem zapytań kierowanych do systemu wyszukiwawczego z wykorzystaniem tej reprezentacji wiedzy jest model wektorowy (*vector space model*)[17].

Bardziej złożone struktury reprezentacji wiedzy to: słownik definicyjny (glosariusz), słownik dziedzinowy, sieć semantyczna i ontologia. Struktury te wprowadzają różne relacje leksykalne występujące pomiędzy przechowywanymi w nich konceptami.

Sieć semantyczna jest grafem skierowanym mającym koncepty (pojęcia) jako wierzchołki oraz krawędzie dla reprezentowania relacji leksykalnych między konceptami. Jest najlepszą strukturą reprezentacji wiedzy do odzwierciedlania powiązań semantycznych między konceptami w języku naturalnym[18], gdyż gromadzi całą wiedzę o semantyce pojęć ze względu na przechowywanie wszystkich relacji leksykalnych charakterystycznych dla języka naturalnego oraz brak nadmiernej złożoności (charakterystycznej dla ontologii).

Stąd wynika jej przydatność w systemach przetwarzających język naturalny. Wnioskowanie z wykorzystaniem sieci semantycznej odbywa się po krawędziach, które mogą posiadać wagi określające ich ważność. Wnioskowanie polega na przeszukiwaniu grafu, w którym, rozpoczynając poruszanie się od jednego węzła grafu (konceptu) i poruszając się po krawędziach (relacje między konceptami) wychodzących z węzła, docieramy do kolejnych węzłów, co odpowiada wnioskowaniu o właściwościach konceptów.

Korzyści wynikające ze stosowania sieci semantycznych w systemach przetwarzających język naturalny zostały opisane przez R.A. Baeza-Yatesa i B. Ribeiro-Neto[19]. Sieci umożliwiają przede wszystkim dostarczenie właściwych znaczeń pojęć, co skutkuje zwiększeniem precyzji odpowiedzi oraz wzrostem pełności odpowiedzi systemu. W zadaniach klasyfikacyjnych ta forma reprezentacji wiedzy podnosi jakość klasyfikacji i kategoryzacji[20].

---

[16] Kolokacja to związek semantyczny, którego znaczenie wynika z połączenia znaczeń kilku słów wchodzących w jego skład (np. „związek małżeński").

[17] R.A. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley Longman Publishing, Boston 1999.

[18] S. Staab, A. Hotho, *Ontology-based text document clustering*, w: *IIS, Advances in Soft Computing*, red. M.A. Kłopotek, S.T. Wierzchoń, K. Trojanowski, Springer, 2003, s. 451-452.

[19] R.A. Baeza-Yates, B. Ribeiro-Neto, op. cit.

[20] M. Baziz, *Towards a Semantic Representation of Documents by Ontology-Document Mapping*, w: *Artificial Intelligence: Methodology, Systems, and Applications. 11th International Confer-*

Powszechnie stosowaną siecią semantyczną dla języka angielskiego jest WordNet[21]. Dla języka polskiego autor posłużył się zbudowaną w ramach projektu SeNeCa[22] siecią semantyczną o nazwie SenecaNet. Sieć ta jest od wielu lat rozbudowywana, a proces rozbudowy obejmuje dodawanie nowych konceptów oraz relacji leksykalnych pomiędzy konceptami. Pierwsze automatyczne metody rozbudowy sieci SenecaNet zaproponowali Ceglarek i Rutkowski[23]. Procedurę jej półautomatycznej rozbudowy z wykorzystaniem kompresji semantycznej opisano w rozdziale 4.1[24].

2.2. Sieć semantyczna SenecaNet

SenecaNet jest siecią semantyczną, która zawiera ponad 154 tysiące konceptów oraz przechowuje rozmaite relacje leksykalne pomiędzy konceptami dla języka polskiego (tab. 1). Sieć ta była pierwszą siecią semantyczną użytą do kompresji semantycznej, m.in. dzięki kilku jej specyficznym własnościom. Koncepty w tej sieci są przechowywane w postaci posortowanej topologicznie listy, co oznacza, że w definicji danego konceptu można odwołać się wyłącznie do konceptów wcześniej zdefiniowanych. W zbudowanej w ten sposób strukturze niemożliwe jest istnienie cykli, dlatego też algorytmy grafowe są wydajniejsze.

Przechowywanie definicji pojęć w postaci listy wynika z szeregu działań optymalizacyjnych, które są stosowane dla szybkiego jej przetwarzania. W tej formie sieć semantyczna może być traktowana jako struktura hierarchiczna, co jest znakomitym rozwiązaniem z obliczeniowego punktu widzenia. Definicja każdego konceptu w sieci SenecaNet zawiera deskryptor konceptu, listę możliwych form morfologicznych, a także hiperonimy konceptu[25], jego synonimy, holonimy, meronimy, konotacje oraz związane z konceptem relacje nienazwane. Notację w formacie sieci semantycznej SenecaNet ilustruje tabela 2.

---

ence, AIMSA 2004, Varna, Bulgaria, September 2-4, 2004. Proceedings, red. Ch. Bussler, D. Fensel, Springer, Berlin – Heidelberg 2004, „Lecture Notes in Computer Science" 2004, t. 3192, s. 33-43.

[21] Projekt Cognitive Science Laboratory Uniwersytetu Princeton jest dostępny pod adresem: http://wordnet.princeton.edu.

[22] Projekt SeNeCa (**Se**mantic **Ne**twork and **Ca**tegorization, http://seneca.kie.ae.poznan.pl) miał za zadanie automatyzację rozbudowy sieci semantycznej dla języka polskiego.

[23] D. Ceglarek, *Zastosowanie sieci semantycznej...*

[24] Szczegółowy opis w: D. Ceglarek, K. Haniewicz, W. Rutkowski, *Towards Knowledge Acquisition…*

[25] Każdy koncept może posiadać jeden lub więcej hiperonimów, dzięki czemu uzyskana struktura jest pod względem taksonomicznym heterarchią; zob. S. Staab, A. Hotho, op. cit.

Tabela 1. Porównanie sieci WordNet i SenecaNet

Parametry	WordNet	Sieć SenecaNet
Liczba konceptów	155 200	154 200
Liczba słów polisemicznych	27 000	21 300
Liczba synonimów		8400
Relacje hiperonimii, hiponimii	+	+
Relacje antonimii	+	–
Konotacje	+	+
Relacje nienazwane	–	+

Ź r ó d ł o: opracowanie własne.

Tabela 2. Format definicji pojęć w sieci SenecaNet

samochód → pojazd, &silnik
Chiny → kraj, :Azja
prowincja.n.01 → jednostka podziału administracyjnego
prowincja.n.02 → obszar geograficzny,*zacofany
prowincja → prowincja.n.01; prowincja.n.02
prowincja Guangdong → prowincja.n.01, :Chiny,
Guangzhou → miasto, :prowincja Guangdong, #stolica(prowincja Guangdong)
Canton → =Guangzhou, *starodawna nazwa chińska

Ź r ó d ł o: opracowanie własne.

2.3. Konwersja sieci semantycznej WordNet

Zastosowanie kompresji semantycznej dla języka angielskiego wymagało użycia sieci semantycznej o strukturze analogicznej do tej, którą posiada sieć semantyczna SenecaNet. Ze względu na trudność zadania utworzenia nowej sieci semantycznej dla języka angielskiego zdecydowano się wykorzystać istniejącą sieć semantyczną WordNet i poddać ją przekształceniu do takiej samej struktury reprezentacji pojęć jak w sieci SenecaNet[26]. Sieć WordNet jest bardo obszerną i dojrzałą siecią semantyczną, która osiągnęła swój obecny kształt dzięki wieloletniej pracy ogromnego zespołu ludzi, i jej przydatność została wykazana w szeregu badań i eksperymentów[27].

[26] Zadanie to zostało szczegółowo opisane w: D. Ceglarek, K. Haniewicz, W. Rutkowski, *Quality of Semantic Compression*…

[27] W pracy J. Rosenzweig, R. Mihalcea, A. Csomai, „*WordNet bibliography". Web page: a bibliography referring to research involving the WordNet lexical database*, http://lit.csci.unt.

Jednakże przekształcenie zorientowanej na synsety struktury sieci Word-Net w nieposiadającą cykli (w rozumieniu grafowym) strukturę operującą deskryptorami konceptów identyfikowanych w analizowanych tekstach okazało się zadaniem trudnym. Dlatego też został opracowany algorytm umożliwiający to przekształcenie, posługujący się zbiorami i bierze pod uwagę każdą lemmę przechowywaną w danym synsecie oraz synsety, które stanowią hiperonimy w stosunku do przetwarzanego synsetu.

Synset definiuje się jako grupę lemm (termów) mających takie samo znaczenie. Po dokładnym przestudiowaniu okazało się, że większość lemm zgromadzonych w jednym synsecie nie stanowi w stosunku do siebie idealnych synonimów. Mają one wspólne znaczenie, lecz poziom podobieństwa znaczeniowego jest różny. Każda lemma występująca w synsecie może być pojedynczym słowem lub związkiem semantycznym, który składa się z kilku słów[28].

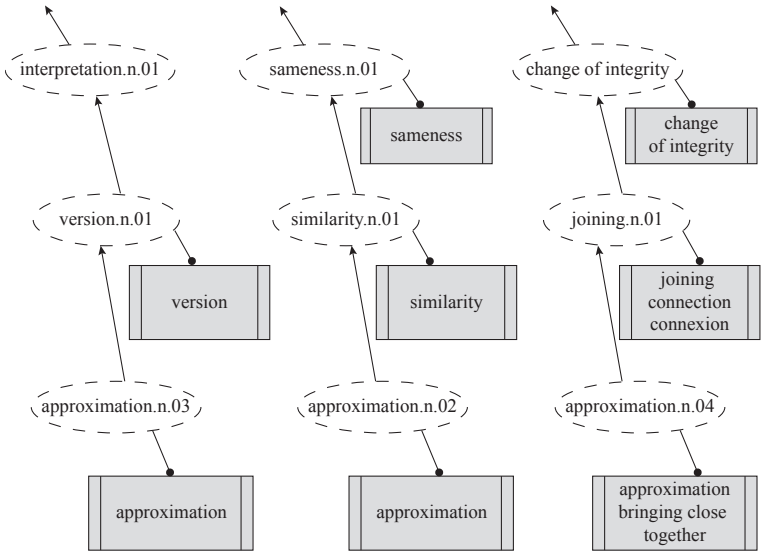
Podczas przekształcania struktury sieci semantycznej ze zorientowanej na synsety w strukturę w pełni hierarchiczną niezbędne jest zmodyfikowanie sposobu wyboru konceptów opisujących dany synset – w celu uniknięcia występowania cykli w docelowym grafie konceptów. Najprostsza sytuacja występuje wtedy, kiedy lemma zawarta w deskryptorze synsetu występuje wyłącznie w tym synsecie, czyli lemma jest unikalnym deskryptorem synsetu (warunek unikalności). W innych przypadkach należy znaleźć inną lemmę z tego samego synsetu, która spełnia warunek unikalności. Przeprowadzone eksperymenty dotyczące rzeczowników wykazały, że postępowanie takie prowadzi do uzyskania sieci semantycznej, w której jest jedynie 25 000 konceptów z 86 000 istniejących w zawierających rzeczowniki synsetach WordNetu.

Wprowadzony został zatem „syntetyczny" deskryptor synsetów, który wynika z potrzeby uniknięcia występowania cykli.

Na rysunkach 1 i 2 przedstawiono wizualizację powyższego procesu przekształcenia na przykładzie lemmy „approximation", która występuje w kilku różnych synsetach. Z tego powodu nie może ona być deskryptorem synsetu. Na rysunku 2 można z łatwością zauważyć, że lemma „bringing close together" występuje dokładnie w jednym synsecie, a zatem może ona zastąpić syntetyczny deskryptor „approximation.n.04" (spełniając warunek unikalności). Procedurę przekształcenia struktury sieci WordNet do formatu SenecaNet opisano poniżej i pokazano w postaci pseudokodu w Algorytmie 1.

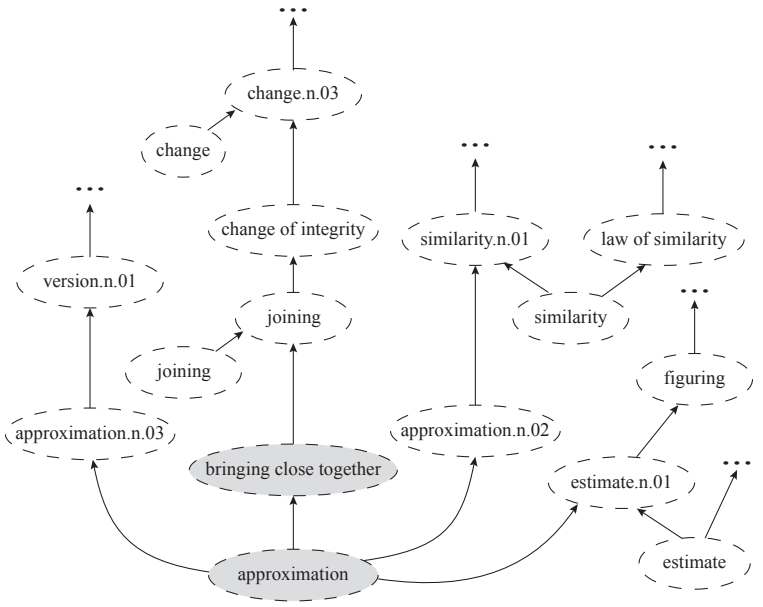
edu/%7Ewordnet [1.09.2007], przedstawiono 868 projektów wykorzystujących WordNet w zadaniach przetwarzania języka naturalnego.

[28] G.A. Miller, *Wordnet: a lexical database for English*, „Communications of the ACM" 1995, t. 38, nr 11.



Rys. 1. Struktura sieci WordNet zorientowana na synsety

Ź r ó d ł o: D. Ceglarek, K. Haniewicz, W. Rutkowski, *Quality of Semantic Compression in Classification*, w: *Computational Collective Intelligence, Second International Conference, ICCCI 2010, Kaohsiung, Taiwan, November 10-12, 2010. Proceedings*, cz. 1, red. J.-S. Pan, S.-M. Chen, N.T. Nguyen, Springer-Verlag, Berlin – Heidelberg 2010, „Lecture Notes in Computer Science" 2010, t. 6421, s. 162-171.



Rys. 2. Taksonomia konceptów w sieci SenecaNet

Ź r ó d ł o: jak przy rys. 1.

Algorytm 1. Algorytm przekształcenia sieci WordNet do formatu SenecaNet (efektem przekształcenia jest sieć WiSENet)

```
for all (d,S) ∈ WN do
  for all l ∈ S do
    F[l]++
  end for
end for
for all (d,S) ∈ WN do
  parsuj lemat z deskryptora w synsecie
  l ← split(d,".")[0]
  if F[l] = 1 then
    lemat może być użyty jako deskryptor synsetu
    d ← l
  else
    for all l ∈ S do
      if F[l] = 1 then
        d ← l
        exit
      end if
    end for
  end if
  SN[d] ← S
end for
```

WN – sieć WordNet w postaci listy synsetów identyfikowanych poprzez deskryptory *d*
S – synset zawierający wiele lemm *l*
F[*l*] – liczba synsetów zawierających lemmę *l*
SN – wynikająca z przekształcenia sieć semantyczna WiSENet

Pierwszym krokiem algorytmu jest zbudowanie słownika frekwencyjnego (F) dla lemm, który zawierać będzie liczbę synsetów zawierających daną lemmę. W tym celu algorytm rozpatruje i sumuje wszystkie synsety w sieci WordNet oraz wszystkie lemmy w synsetach. W następnym kroku algorytm pobiera deskryptor (w miarę możliwości lemmę) dla każdego synsetu. Następnie algorytm sprawdza, czy taka lemma występuje dokładnie w jednym synsecie – i jeśli odpowiedź jest pozytywna, to pobrana lemma może być użyta jako nowy deskryptor synsetu. W przeciwnym wypadku sprawdza pozostałe lemmy z analizowanego synsetu i sprawdza, czy istnieje taka, którą można wykorzystać jako deskryptor synsetu. Jeśli nie istnieje wśród nich żadna lemma spełniająca warunek unikalności, jako deskryptor synsetu zostaje wykorzystany oryginalny deskryptor z sieci WordNet.

Uzyskana w wyniku powyższego przekształcenia sieć semantyczna dla języka angielskiego WiSENet zawiera te same dane (koncepty i relacje leksykalne), które zawiera sieć semantyczna WordNet, jednakże ma hierarchiczną strukturę sieci SenecaNet.

3. Kompresja semantyczna

Zgodnie z definicją podaną na wstępie kompresja semantyczna jest techniką, która ma za zadanie dostarczyć bardziej ogólne koncepty w stosunku do konceptu, który wystąpił w analizowanym dokumencie lub w zapytaniu kierowanym do systemu. Konceptu istniejącego w danym dokumencie w pewnym kontekście, który decyduje o jego znaczeniu.

Gdy zadaniem algorytmu kompresji semantycznej jest wyznaczenie konceptu bardziej generalnego w stosunku do danego konceptu, algorytm musi precyzyjnie określić poziom tej generalizacji. Im wyższy jest poziom generalizacji, tym większa jest utrata informacji. W niektórych zastosowaniach może to stanowić pozytywne zjawisko (np. w zadaniu klasyfikacji dokumentów metodami analizy skupień[29]), ale wtedy, kiedy kompresja semantyczna ma służyć przekształceniu dokumentu w celu przedstawienia go odbiorcy będącemu człowiekiem, nie jest to zjawisko akceptowalne. W kompresji chodzi o wyznaczenie dla każdego pojęcia takiego pojęcia, które będzie jego reprezentantem (deskryptorem). Podstawą będzie liczba wystąpień w korpusie dokumentów (pojęcia częste będą reprezentowane przez same siebie, pozostałe pojęcia będą reprezentowane przez pojęcia nadrzędne w strukturze, których skumulowana liczba wystąpień jest duża).

Kompresja semantyczna powstała pierwotnie jako kompresja globalna. W związku z pojawieniem się licznych jej zastosowań i wykorzystaniem korpusów dokumentów z rozmaitych dziedzin została rozwinięta poprzez doskonalenie strategii generalizacji w zależności od dziedziny dokumentów, czego efektem jest kompresja dziedzinowa. Prezentuje to następna sekcja, w której zostały przedstawione również wyniki eksperymentów obrazujących skuteczność i efektywność kompresji.

3.1. Mechanizm kompresji semantycznej

Mechanizm kompresji semantycznej został zaprezentowany w 2010 r.[30] jako metoda podnosząca jakość i efektywność klasyfikacji dokumentów tekstowych.

---

[29] R. Nock, F. Nielsen, *On weighting clustering*, „The IEEE Transactions on Pattern Analysis and Machine Intelligence" 2006, nr 28(8), s. 1223-1235.

[30] Zob. D. Ceglarek, K. Haniewicz, W. Rutkowski, *Semantic Compression…*

Kompresja tekstu jest możliwa poprzez zastosowanie sieci semantycznej oraz danych o częstości wystąpień konceptów (w formie słowników frekwencyjnych). Efektem takiego postępowania jest redukcja liczby konceptów używanych do reprezentowania pojęć występujących w dokumentach tekstowych bez znaczącej straty informacji, co jest niezwykle istotne z perspektywy procesu przetwarzania języka naturalnego (zwłaszcza wtedy, kiedy stosuje się model wektorowy[31]).

Ponadto redukcja liczby konceptów pomaga w radzeniu sobie ze zjawiskami lingwistycznymi, które stanowią znaczne wyzwanie w zadaniach przetwarzania języka naturalnego[32]. Zjawiskiem, na które najczęściej zwraca się uwagę w zadaniach NLP, jest polisemia oraz synonimia[33]. Gdy używa się wielu termów jako określenia tego samego lub podobnego konceptu, to mogą one zostać zastąpione jednym, bardziej ogólnym konceptem. Przy zastosowaniu analizy statystycznej można sporządzić słownik frekwencyjny dopasowany do kontekstu danej dziedziny i wyznaczyć właściwy deskryptor dla konceptów polisemicznych. Uzyskana w wyniku zredukowania zbioru konceptów struktura jest wydajniejsza obliczeniowo i powoduje mniejszą utratę informacji niż rozwiązania, które nie stosują tej techniki.

Przyjmijmy, że posiadamy dokumenty poświęcone badaniom biologicznym i w związku z tym w dokumentach występują łacińskie nazwy gatunków zwierząt lub roślin. Podczas klasyfikacji takich dokumentów te łacińskie nazwy występujące w dokumentach poszerzają wektor termów opisujący dokumenty, co utrudnia obliczeniowo proces ich klasyfikacji oraz powoduje spadek jakości uzyskanej klasyfikacji. Zamiana nazwy łacińskiej określającej gatunek jakiejś rośliny na odpowiadający jej koncept skraca wynikowy wektor pojęć opisujących dokument oraz skutkuje minimalną stratą informacyjną. Naturalnie, taka zamiana może być dokonana dla specyficznego korpusu dokumentów, w którym nazwy łacińskie są stosunkowo rzadkie, a zatem możliwe do zastąpienia. Wybór konceptów w procesie generalizacji pojęć jest zależny od dziedziny dokumentów.

W ogólności, kompresja semantyczna umożliwia realizowanie zadań przetwarzania języka naturalnego, takich jak wyszukiwanie wzorców w tekstach, operując na poziomie konceptów, a nie pojedynczych słów. Osiąga się to nie tylko przez reprezentowanie termów przez ich wspólne znaczenie (rozwiązanie znane jako podejście zorientowane na synsety[34]), ale również poprzez zastępowanie długich fraz ich krótszymi odpowiednikami.

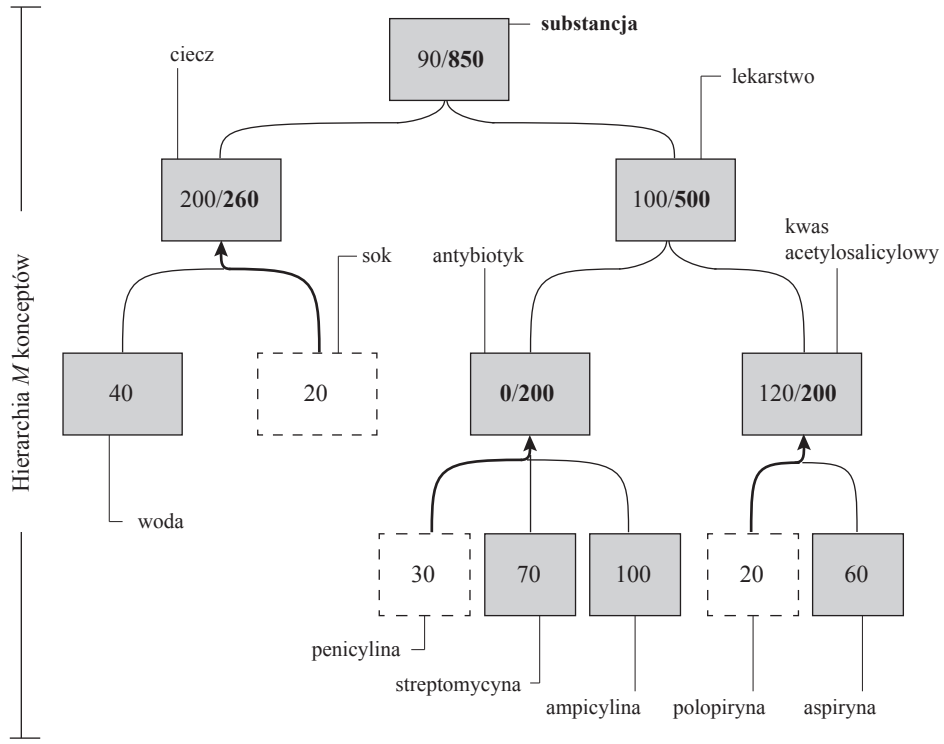
---

[31] R.A. Baeza-Yates, B. Ribeiro-Neto, op. cit.; K. Erk, S. Pad'o, *A Structured Vector Space Model for Word Meaning in Context*, w: *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA 2008, s. 897-906.

[32] R. Sinha, R. Mihalcea, *Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity*, w: *International Conference on Semantic Computing ICSC 2007*, IEEE 2007, s. 363-369.

[33] R. Krovetz, W.B. Croft, *Lexical ambiguity and information retrieval*, „ACM Transactions on Information Systems" 1992, nr 10, s. 115-141.

[34] G.A. Miller, op. cit.



Rys. 3. Wybór konceptów o największej skumulowanej liczbie wystąpień spośród *M* konceptów w sieci semantycznej

Ź r ó d ł o: opracowanie własne.

Kompresja semantyczna pozwala na znajdowanie wspólnego znaczenia dla sentencji wyrażonych za pomocą różnych termów.

Mechanizm uogólniania pojęć nie jest zdolny do analizy zależności pomiędzy nimi oraz do dokonywania zmian w sentencjach zgodnie z zasadami gramatyki. Umożliwia jednak trafne wykrywanie, że różnie sformułowane sentencje niosą tę samą zawartość informacyjną. W zadaniu wykrywania plagiaryzmów w ramach omawianego systemu SOWI wykorzystywany jest algorytm „*bag of concepts*", którego wyróżniającymi cechami są: niewrażliwość na zmianę szyku termów w porównywanych dokumentach oraz niewrażliwość na stosunkowo krótkie niezgodności w sekwencjach termów[35]. Ceglarek i Haniewicz zaproponowali algorytm o tych samych własnościach, lecz mający znacznie mniejszą, bo logarytmiczną złożoność obliczeniową[36].

[35] Szczegółowy opis w: D. Ceglarek, *Koncepcja komponentowego systemu ochrony…*

[36] D. Ceglarek, K. Haniewicz, *Fast Plagiarism Detection by Sentence Hashing*, w: *Artificial Intelligence and Soft Computing. 11th International Conference, ICAISC 2012, Zakopane, Poland,*

Należy rozumieć, że kompresja semantyczna jest mechanizmem, którego zastosowanie oznacza utratę części informacji semantycznej, lecz utrata informacji jest nieznaczna, kiedy wybrane deskryptory pojęć bardziej generalnych są konceptami często występującymi w dokumentach tekstowych i znaczenie wybranych konceptów bardziej ogólnych jest podobne. Poziomem kompresji steruje się poprzez określenie liczby konceptów, które stają się deskryptorami używanymi do opisu tekstu dokumentów. Eksperymenty przeprowadzone w celu zmierzenia jakości metod w ramach zadań przetwarzania języka naturalnego pokazały, że redukcja liczby konceptów do ok. 4000 nie wpływa znacząco na utratę jakości w zadaniach klasyfikacji dokumentów. Idea wyboru konceptów zgodnie z ich częstością występowania w dokumentach pokazana została na rysunku 3. Opis samego algorytmu znajduje się w następnym punkcie artykułu.

3.2. Algorytm kompresji semantycznej

W korpusie dokumentów występuje M konceptów k_i, które można użyć do utworzenia M-elementowego wektora reprezentującego dokumenty. Określona jest również docelowa liczba konceptów N (gdzie $N < M$). W pierwszej kolejności należy obliczyć liczbę wystąpień $f(k_i)$ dla każdego konceptu k_i we wszystkich dokumentach. Następnie należy obliczyć skumulowaną liczbę wystąpień dla wszystkich konceptów – do liczby wystąpień danego konceptu dodaje się sumę wystąpień wszystkich jego hiponimów. Kolejnym krokiem jest włączenie informacji o liczbie wystąpień wynikających z relacji synonimii w ten sposób, że dla grupy konceptów połączonych relacją synonimii wybiera się koncept mający największą skumulowaną liczbę wystąpień i do jego skumulowanej liczby wystąpień dodaje się skumulowane liczby wystąpień wszystkich pozostałych synonimów.

Poruszając się w górę hierarchii konceptów, należy obliczyć skumulowaną liczbę wystąpień konceptów poprzez dodanie sumy skumulowanej liczby wystąpień hiponimów do danego konceptu (ich hiperonimu):
$cumf(k_i) = f(k_i) + \sum_j \left[(cumf(k_j) \right]$, gdzie k_i jest hiperonimem dla k_j (patrz Algorytm 2 oraz Algorytm 3).

Ostatnim krokiem algorytmu jest wybranie N konceptów o największej skumulowanej liczbie wystąpień, które stanowić będą listę deskryptorów (Algorytm 4). Opisana procedura kompresji semantycznej pozwala zredukować rozmiar wektora pojęć opisujących dokumenty o M-N konceptów.

April 29-May 3, 2012, Proceedings, t. 2, red. L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, J.M. Zurada, Springer-Verlag, Berlin – Heidelberg 2012, „Lecture Notes in Computer Science" 2012, t. 7268, s. 30-38.

Algorytm 2. Ustalenie uogólnionych konceptów, które staną się deskryptorami pojęć poprzez obliczenie skumulowanej liczby wystąpień w korpusie dokumentów C

```
//wybór synonimu reprezentującego grupę pojęć synonimicznych
max = 0
n = 0
sum = 0
for s ∈ S_v do
  sum = sum + l_s
  if l_s > max then
    max = l_s
    n = s
  end if
end for
l_s = l_s + sum
```

```
// obliczenie skumulowanej liczby wystąpień dla hiperonimów
for v ∈ V'' do
  p = card(H_v)
  for h ∈ H_v do
    l_h = l_h + l_v / p
  end for
end for
```

S_v – zbiór synonimów dla konceptu v
V – wektor konceptów przechowywany w sieci semantycznej
V' – topologicznie posortowany wektor V
V'' – odwrócony wektor V'
l_v – liczba wystąpień konceptu v w korpusie dokumentów C
H_v – zbiór hiperonimów konceptu v

Algorytm 3. Wybór m konceptów z sieci semantycznej w procedurze dziedzinowej kompresji semantycznej

```
for v ∈ V do
  if l_v ≥ f
    d_v = v
  else
    d_v = FMax(v)
  end if
end for
```

L – wektor przechowujący liczbę wystąpień konceptów w korpusie dokumentów C
L' – posortowany malejąco wektor L
f – liczba wystąpień m-tego konceptu w wektorze L'

Algorytm 4. Procedura FMax znajdująca dla danego konceptu *v* jego deskryptor (hiperonim o największej skumulowanej liczbie wystąpień)

FMax(*v*):
$max = 0$
$x = \varnothing$
for $h \in H_v$ **do**
  if $d_h \neq \varnothing$ **then**
    if $l_{d_h} > max$ **then**
      $max = l_{d_h}$
      $x = d_h$
    end if
  end if
end for
return *x*

W tabeli 3 zamieszczone zostały dwa przykłady semantycznie skompresowanych fragmentów tekstu (przy 4000 deskryptorach konceptów) w języku angielskim. Po oryginalnych fragmentach (A, B) przytoczone są fragmenty skompresowane (A', B').

Tabela 3. Przykłady kompresji semantycznej dla języka angielskiego

A	The information from AgCam will provide useful data to agricultural producers in North Dakota and neighboring states, benefiting farmers and ranchers and providing ways for them to protect the environment.
A'	information will provide useful data economic producer american state adjective state benefit creator creator provide structure protect environment
B	Together the two groups make up nearly 70 percent of all flowering plants and are part of a larger clade known as Pentapetalae, which means five petals. Understanding how these plants are related is a large undertaking that could help ecologists better understand which species are more vulnerable to environmental factors such as climate change.
B'	together two group constitute percent group flowering plant part flowering plant known means five leafage understanding plant related large undertaking can help biologist better understand species more sensitive environmental factor such climate change.

Źródło: opracowanie własne.

3.3. Ocena jakości kompresji semantycznej

Przyjmijmy, że zadaniem systemu jest przetworzenie niezbyt specjalistycznego artykułu poświęconego najnowszym osiągnięciom w rozwoju antybiotyków. Po ustaleniu dziedziny dokumentu, co jest stosunkowo prostym zadaniem w ramach obecnie dostępnych systemów klasyfikacyjnych, można przystąpić do kompresji z wykorzystaniem sieci semantycznej. Przetwarzając ów artykuł, można zauważyć, że każde odniesienie do penicyliny lub streptomycyny jest miejscem stosownym do zastosowania kompresji. Sieć semantyczna zawiera relacje pozwalające na wywnioskowanie, że zarówno penicylina, jak i streptomycyna są antybiotykami. Rezultatem zastosowania kompresji jest skrócenie wektora opisującego dokument o dwa elementy poprzez zastąpienie konkretnych nazw antybiotyków ich generalizacją. Analogiczny proces może być zastosowany do kolejnych termów, których poziom specjalizacji odbiega od średniego poziomu artykułu.

Przygotowany został eksperyment mający określić, czy kompresję semantyczną można skutecznie zastosować w typowym dla *text miningu* zadaniu klasyfikacji dokumentów (zastosowano klasyfikację aglomeracyjną metodą Warda)[37]. W pierwszym przebiegu dokumenty klasyfikowane były w oryginalnej postaci, a w drugim przebiegu zostały poddane kompresji semantycznej. Do eksperymentu posłużyło 900 dokumentów tekstowych w języku angielskim z zakresu astronomii, biologii, ekonomii, kultury, medycyny, polityki, prawa oraz sportu. W celu zweryfikowania rezultatów wszystkie dokumenty zostały zaetykietowane manualnie listą kategorii, do których zostały zaliczone przez eksperta.

Klasyfikację metodą analizy skupień przeprowadzono ośmiokrotnie. Pierwszy przebieg został przeprowadzony bez zastosowania kompresji semantycznej: z użyciem ok. 25 000 konceptów. W następnych przebiegach algorytmu redukowano liczbę konceptów będących deskryptorami do 12 000, 10 000, 8000, 6000, 4000, 2000 – aż do kompresji semantycznej z użyciem 1000 deskryptorów konceptów.

Jakość klasyfikacji została obliczona poprzez porównanie dokonanej klasyfikacji z etykietami nadanymi dokumentom przez ekspertów. Uzyskane współczynniki jakości klasyfikacji zostały zaprezentowane w tabelach 4 oraz 5. W wynikach można zaobserwować, że nieznaczny spadek jakości klasyfikacji występuje dla silnej kompresji semantycznej (liczba konceptów zostaje zredukowana poniżej 4000).

---

[37] A. Hotho, S. Staab, G. Stumme, *Explaining Text Clustering Results Using Semantic Structures*, w: *Knowledge Discovery in Databases: PKDD 2003. 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings*, red. N. Lavrač, D. Gamberger, H. Blockeel, L. Todorovski, PKDD, Springer Verlag, Berlin – Heidelberg 2003, „Lecture Notes in Computer Science" 2003, t. 2838, s. 217-228.

Tabela 4. Oszacowanie jakości klasyfikacji dokumentów oryginalnych (bez kompresji semantycznej) w proc.

Liczba cech / Liczba konceptów	1000	900	800	700	600	Średnia
bez kompresji	93,46	90,90	91,92	92,69	89,49	91,69
12 000 konceptów	91,92	90,38	90,77	88,59	87,95	89,92
10 000 konceptów	93,08	89,62	91,67	90,51	90,90	91,15
8000 konceptów	92,05	92,69	90,51	91,03	89,23	91,10
6000 konceptów	91,79	90,77	90,90	89,74	91,03	90,85
4000 konceptów	88,33	89,62	87,69	86,79	86,92	87,87
2000 konceptów	86,54	87,18	85,77	85,13	84,74	85,87
1000 konceptów	83,85	84,10	81,92	81,28	80,51	82,33

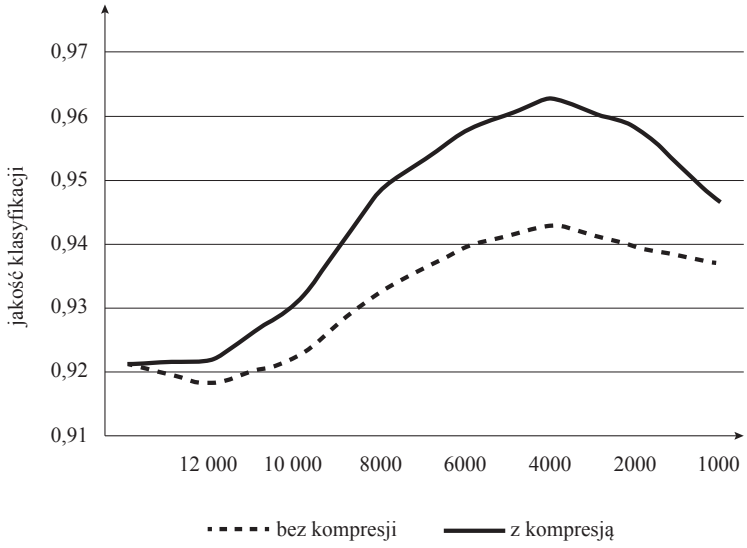
Ź r ó d ł o: opracowanie własne.

Tabela 5. Jakość klasyfikacji dokumentów przy zastosowaniu kompresji semantycznej w proc.

Liczba cech / Liczba konceptów	1000	900	800	700	600	Średnia
wszystkie koncepty	94,78	92,50	93,22	91,78	91,44	92,11
12 000 konceptów	93,56	93,39	93,89	91,50	91,78	92,20
10 000 konceptów	95,72	94,78	93,89	91,61	92,17	93,08
8000 konceptów	95,89	95,83	94,61	95,28	94,72	94,86
6000 konceptów	96,94	96,11	96,28	96,17	95,06	95,77
4000 konceptów	96,83	96,33	96,89	96,06	96,72	96,27
2000 konceptów	97,06	96,28	95,83	96,11	95,56	95,83
1000 konceptów	96,22	95,56	94,78	94,89	94,00	94,66

Ź r ó d ł o: opracowanie własne.

Rysunek 1 pokazuje jakość klasyfikacji uzyskanej w obu zadaniach. Utrata jakości klasyfikacji dokumentów jest nieznaczna dla kompresji, która redukuje liczbę deskryptorów konceptów do 4000. Natomiast silniejsza kompresja i związana z nią mniejsza liczba deskryptorów powoduje znaczne zmniejszenie jakości klasyfikacji, która jednak pozostaje na akceptowalnym poziomie.



Rys. 1. Jakość klasyfikacji dokumentów określająca w procentach prawidłowo sklasyfikowane dokumenty

Ź r ó d ł o: opracowanie własne.

4. Zastosowania kompresji semantycznej

Jednym z najważniejszych zastosowań kompresji semantycznej okazała się metoda półautomatycznej rozbudowy samej sieci. Inne zastosowanie polega na takim zaprezentowaniu użytkownikowi dokumentów w sposób uogólniony, aby dopasować uzyskane w wyniku uogólnienia pojęcia do stopnia kompetencji użytkownika w dziedzinie, której dotyczy dokument.

4.1. Rozbudowa sieci semantycznej. Algorytm regułowego wykrywania pojęć i relacji leksykalnych

Do wykrywania nowych pojęć skonstruowano algorytm, który został następnie użyty w eksperymencie z wykorzystaniem sieci semantycznej WiSENet. Pierwszym krokiem w algorytmie jest procedura rozwijająca reguły w zbiory hiponimów zawartych w sieci. Operacja ta jest kosztowna czasowo ze względu na konieczność przejścia po wszystkich możliwych krawędziach łączących wybrane koncepty oraz wszystkich konceptach końcowych w użytej sieci semantycznej.

Następnym etapem jest przeanalizowanie tekstów pod kątem pasujących fragmentów w tekstach. Operacja ta jest wykonywana z wykorzystaniem mechanizmu o nazwie *bag of concepts*, zaimplementowanego jako automat skończony wyposażony w zaawansowane metody wyzwalające zaprogramowane operacje. W każdym stanie automatu sprawdza on, czy którakolwiek z reguł wymagających sprawdzenia jest spełniona. Kompletny opis algorytmu znajduje się w pracy Ceglarka i wsp.[38], w pseudokodzie pokazany jest w Algorytmie 5.

Tabela 6. Przykład reguły i rezultatów działania automatu *bag of concepts*

Reguła: disease (wszystkie hiponimy), therapy (wszystkie hiponimy)
Znaleziony fragment: chemotherapy drug finish off remaining cancer **Koncepty spełniające regułę**: therapy → chemotherapy, disease → cancer **Ignorowane**: drug finish off remaining
Znaleziony fragment: gene therapy development lymphoma say woods **Koncepty spełniające regułę**: therapy → gene therapy, disease → lymphoma **Ignorowane**: development
Znaleziony fragment*: cancer by-bid using surgery chemotherapy* ***Koncepty spełniające regułę****: therapy → chemotherapy, disease → cancer* ***Ignorowane****: by-bid using surgery*

Ź r ó d ł o: opracowanie własne.

Podany w tabeli 6 przykład pochodzi z eksperymentów przeprowadzonych na korpusie 2589 angielskich tekstów z dziedziny biologii i medycyny (łącznie dokumenty zawierały ponad 9 milionów słów). Rezultatem eksperymentu było znalezienie 471 nowych pojęć, które zostały następnie dodane do sieci WiSENet.

4.2. Mechanizm wspomagania rozumienia tekstu

Dziedzinowa kompresja semantyczna została sprawdzona również w zastosowaniu społecznościowym. Dla dokumentów w języku polskim (z dziedziny astronomii, biologii oraz astrobiologii) przeprowadzony został eksperyment z użyciem sieci semantycznej SenecaNet wraz z dodatkowym mechanizmem wykorzystującym analizator morfologiczny Morfologik[39]. Eksperyment polegał na dostosowaniu siły kompresji semantycznej do potrzeb użytkownika (w zależności od deklarowanego stopnia posiadanych kompetencji w danej dziedzinie). Jednocześnie zadaniem systemu było zaprezentowanie przekształconego tekstu w formie

[38] D. Ceglarek, K. Haniewicz, W. Rutkowski, *Towards Knowledge Acquisition…*

[39] M. Miłkowski, op. cit.

Algorytm 5. Algorytm automatu skończonego *bag of concepts* do wykrywania reguł z użyciem sieci semantycznej WiSENet

```
//przypisanie wyzwalaczy reguł do konceptów w sieci semantycznej
mapRulesToSemNet(SN, R[])
for all Rule ∈ R do
  for all Term, Relations ∈ Rule do
    N = SN.getNeighbourhood(Term, Relations)
    for all Term ∈ N do
      SN.createRuleTrigger(Term, Rule)
    end for
  end for
end for
//wygładzenie tekstu: tokenizacja, zastosowanie stop-listy, wykrywanie pojęć.
T = analyzeText(Input)
for each Term ∈ T
  if count(Bag) = size(Bag) then
    //deaktywowanie licznika wystąpień dla reguł związanych z termem Term.
    //wyjęcie termu ze zbioru Bag.
    oldTerm = pop(Bag)
  end if
  for all Rule ∈ SN.getTriggers(oldTerm) do
    Rule.unhit(Term)
    push(Bag, Term)
    for all Rule ∈ SN.getTriggers(Term) do
      //pobranie relewantnych reguł i aktywowanie licznika wystąpień termu.
      Rule.hit(Term)
      if Rule.hitCount = Rule.hitRequired then
        //wyświetlenie raportu informującego o spełnieniu reguły Rule
        Report(Rule, Bag)
      end if
  end for
end for
```

SN – sieć semantyczna WiSENet
R – zbiór reguł semantycznych
Bag – zbiór termów aktywowanych za pomocą automatu skończonego

Tabela 7. Przykład oryginalnego i skompresowanego fragmentu tekstu w języku polskim z wykorzystaniem analizatora morfologicznego Morfologik

Tekst oryginalny: „Zaćmienie księżyca"
O godzinie 19:42:06 Księżyc dotknie cienia Ziemi. Stopniowo od wschodniej strony nasz satelita będzie „pożerany" przez cień naszej planety. O godzinie 20:49:34 cień całkowicie pochłonie Księżyc. Jego barwa powinna stać się krwisto czerwona na skutek oświetlenia promieniami słonecznymi zagiętymi w ziemskiej atmosferze. Maksimum zaćmienia wypadnie o godzinie 21:20:36.
Tekst skompresowany (4000 deskryptorów dla konceptów z sieci)
O godzinie 19:42:06 Księżyc dotknie cienia Ziemi. Stopniowo od wschodniej strony nasz satelita będzie konsumowany przez cień naszej planety. O godzinie 20:49:34 cień całkowicie przyłączy Księżyc. Jego barwa powinna stać się kolorowo czerwona na skutek działania promieniami słonecznymi nierównymi w ziemskiej atmosferze. Maksimum zaćmienia **usunie** o godzinie 21:20:36.

Ź r ó d ł o : opracowanie własne.

zrozumiałej i poprawnej stylistycznie. Zastosowanie mechanizmu wykorzystującego Morfologik pozwoliło w sposób automatyczny dopasowywać formy deklinacyjne i koniugacyjne termów podlegających kompresji semantycznej[40]. W eksperymencie 95,5% transformacji zostało dokonanych poprawnie, uwzględniając wszelkie aspekty gramatyczne w języku polskim. Przykład ilustrujący uzyskane wyniki eksperymentu pokazany jest w tabeli 7.

5. Podsumowanie

Przeprowadzono szereg badań i eksperymentów, które miały na celu rozwinięcie koncepcji kompresji semantycznej i pokazanie jej rozmaitych zastosowań w dziedzinie przetwarzania języka naturalnego. Wyniki badań pokazały, że kompresja semantyczna może być z powodzeniem używana w rozmaitych zadaniach NLP. W pracy omówione zostały następujące istotne rezultaty przeprowadzonych badań:

– notacja SenecaNet dla sieci semantycznej,
– mechanizm globalnej i dziedzinowej kompresji semantycznej,

[40] Rozwiązanie zostało przedstawione w: D. Ceglarek, K. Haniewicz, W. Rutkowski, *Domain Based Semantic Compression for Automatic Text Comprehension Augmentation and Recommendation*, w: *Computational Collective Intelligence. Technologies and Applications. Third International Conference, ICCCI 2011, Gdynia, Poland, September 21-23, 2011, Proceedings*, t. 2, red. P. Jędrzejowicz, N.T. Nguyen, K. Hoang, Springer-Verlag, Berlin – Heidelberg 2011, „Lecture Notes in Computer Science" 2011, t. 6923, s. 40-49.

– mechanizm transformacji sieci semantycznej WordNet do formatu sieci SenecaNet,

– mechanizm łączący kompresję semantyczną z analizą morfologiczną do wspomagania rozumienia dokumentów w wybranych dziedzinach,

– automat skończony dla wyszukiwania nowych pojęć i nowych relacji leksykalnych.

W wyniku przeprowadzonych eksperymentów pokazano, że jakość klasyfikacji dokumentów z wykorzystaniem kompresji semantycznej wzrasta z 92,11% o dodatkowe 4,16%. Dzięki kompresji semantycznej możliwe stało się zbudowanie mechanizmu posługującego się stosunkowo ogólnymi regułami, które skutecznie wykrywają nowe pojęcia w dokumentach tekstowych. Autor zamierza wyszukać nowe zastosowania dla kompresji semantycznej. Dodatkowym zadaniem badawczym jest też udoskonalenie narzędzi i metod służących do w pełni automatycznej rozbudowy sieci semantycznej WiSENet.

Literatura

Baeza-Yates R.A., Ribeiro-Neto B., *Modern Information Retrieval*, Addison-Wesley Longman Publishing, Boston 1999.

Baziz M., *Towards a Semantic Representation of Documents by Ontology-Document Mapping*, w: *Artificial Intelligence: Methodology, Systems, and Applications. 11th International Conference, AIMSA 2004, Varna, Bulgaria, September 2-4, 2004. Proceedings*, red. Ch. Bussler, D. Fensel, Springer, 2004, „Lecture Notes in Computer Science" 2004, t. 3192, s. 33-43.

Boyd-Graber J., Blei D.M., Zhu X., *A Topic Model for Word Sense Disambiguation*, w: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, June 2007*, s. 1024-1033.

Burrows S., Tahaghoghi S.M.M., Zobel J., *Efficient plagiarism detection for large code repositories*, „Software: Practice and Experience" 2007, t. 37, nr 2, s. 151-175.

Ceglarek D., *Zastosowanie sieci semantycznej do disambiguacji pojęć w języku naturalnym*, w: *Systemy wspomagania organizacji SWO 2006*, Wyd. AE w Katowicach, Katowice 2006.

Ceglarek D., *Koncepcja komponentowego systemu ochrony własności intelektualnej wykorzystującego semantyczne struktury informacji*, w: *Technologie informatyczne w zarządzaniu wiedzą – uwarunkowania i realizacja*, red. P. Adamczewski, M. Zakrzewicz, Wyd. WSB w Poznaniu, Poznań 2009.

Ceglarek D., Haniewicz K., Rutkowski W., *Quality of Semantic Compression in Classification*, w: *Computational Collective Intelligence, Second International Conference, ICCCI 2010, Kaohsiung, Taiwan, November 10-12, 2010. Proceedings*, cz. 1, red. J.-S. Pan, S.-M. Chen, N.T. Nguyen, Springer-Verlag, Berlin – Heidelberg 2010, „Lecture Notes in Computer Science" 2010, t. 6421, s. 162-171.

Ceglarek D., Haniewicz K., Rutkowski W., *Semantic Compression for Specialised Information Retrieval Systems*, w: *Advances in Intelligent Information and Database Systems*, red. N.T. Nguyen, R. Katarzyniak, S.-M. Chen, Springer Verlag, Berlin – Heidelberg 2010, „Studies in Computational Intelligence" 2010, t. 283, s. 111-121.

Ceglarek D., Haniewicz K., Rutkowski W., *Domain Based Semantic Compression for Automatic Text Comprehension Augmentation and Recommendation*, w: *Computational Collective Intelligence. Technologies and Applications. Third International Conference, ICCCI 2011, Gdynia, Poland, September 21-23, 2011, Proceedings*, t. 2, red. P. Jędrzejowicz, N.T. Nguyen, K. Hoang, Springer-Verlag, Berlin – Heidelberg 2011, „Lecture Notes in Computer Science" 2011, t. 6923, s. 40-49.

Ceglarek D., Haniewicz K., Rutkowski W., *Towards Knowledge Acquisition with WiSENet*, w: *New Challenges for Intelligent Information and Database Systems*, red. N.T. Nguyen, B. Trawinski, J.J. Jung, Springer Verlag, Berlin – Heidelberg 2011, „Studies in Computational Intelligence" 2011, t. 351, s. 75-84.

Ceglarek D., Haniewicz K., *Fast Plagiarism Detection by Sentence Hashing*, w: *Artificial Intelligence and Soft Computing. 11th International Conference, ICAISC 2012, Zakopane, Poland, April 29-May 3, 2012, Proceedings*, t. 2, red. L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, J.M. Zurada, Springer-Verlag, Berlin – Heidelberg 2012, „Lecture Notes in Computer Science" 2012, t. 7268, s. 30-38.

Erk K., Pad'o S., *A Structured Vector Space Model for Word Meaning in Context*, w: *EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA 2008, s. 897-906.

Hotho A., Staab S., Stumme G., *Explaining Text Clustering Results Using Semantic Structures*, w: *Knowledge Discovery in Databases: PKDD 2003. 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003*, *Proceedings*, red. N. Lavrač, D. Gamberger, H. Blockeel, L. Todorovski, PKDD, Springer Verlag, Berlin – Heidelberg 2003, „Lecture Notes in Computer Science" 2003, t. 2838, s. 217-228.

Information Retrieval: Data Structures & Algorithms, red. W.B. Frakes, R.A. Baeza-Yates, Prentice-Hall, 1992.

Krovetz R., Croft W.B., *Lexical ambiguity and information retrieval*, „ACM Transactions on Information Systems" 1992, nr 10, s. 115-141.

Lukashenko R., Graudina V., Grundspenkis J., *Computer-based plagiarism detection methods and tools: an overview*, w: *Proceedings of the 2007 International Conference on Computer Systems and Technologies, CompSysTech '07. New York, USA*, ACM, 2007, s. 401-406.

Miller G.A., *Wordnet: a lexical database for English*, „Communications of the ACM" 1995, t. 38, nr 11.

Miłkowski M., *Automated Building of Error Corpora of Polish*, w: *Corpus Linguistics, Computer Tools, and Applications – State of the Art*, PALC 2007, red. B. Lewandowska-Tomaszczyk, Peter Lang, Frankfurt am Main 2008, s. 631-639.

Nock R., Nielsen F., *On weighting clustering*, „The IEEE Transactions on Pattern Analysis and Machine Intelligence" 2006, nr 28(8), s. 1223-1235.

Ota T., Masuyama S., *Automatic plagiarism detection among term papers*, w: *Proceedings of the 3rd International Universal Communication '09*, ACM, 2009, s. 395-399.

Percova N.N., *On the types of semantic compression of text*, w: *COLING '82. Proceedings of the 9th conference on Computational linguistics*, t. 2, Academia Praha, 1982, s. 229-231.

Rosenzweig J., Mihalcea R., Csomai A., „*WordNet bibliography". Web page: a bibliography referring to research involving the WordNet lexical database*, http://lit.csci.unt.edu/%7Ewordnet [1.09.2007].

Sanderson M., *Word Sense Disambiguation and Information Retrieval*, w: *SIGIR '94. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, red. W.B. Croft, C.J. van Rijsbergen, SIGIR, ACM/Springer, New York 1994, s. 142-151.

Sanderson M., *Retrieving with Good Sense*, „Information Retrieval" 2000, t. 2, nr 1, s. 49-69.

Sinha R., Mihalcea R., *Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity*, w: *International Conference on Semantic Computing ICSC 2007*, IEEE 2007, s. 363-369.

Snow R., Jurafsky D., Ng A.Y., *Learning syntactic patterns for automatic hypernym discovery*, w: *Advances in Neural Information Processing Systems (NIPS)*, 2005.

Staab S., Hotho A., *Ontology-based text document clustering*, w: *IIS, Advances in Soft Computing*, red. M.A. Kłopotek, S.T. Wierzchoń, K. Trojanowski, Springer, 2003, s. 451-452.

Stokoe Ch., Oakes M.P., Tait J., *Word Sense Disambiguation in Information Retrieval Revisited*, SIGIR, 2003.