Eksploracja danych reprezentowanych za pomocą języka naturalnego, z wykorzystaniem ontologii

Autoreferat rozprawy doktorskiej

Wojciech Jaworski
Instytut Informatyki
Uniwersytet Warszawski

10 grudnia 2008

1 Sformułowanie problemu

Upowszechnienie komputerowych metod przetwarzania informacji, a zwłaszcza komputeryzacja zasobów wiedzy, doprowadziły do powstania bogatych korpusów tekstów w systemach komputerowych. Aby korzystać z takich zbiorów dokumentów, potrzebne są metody wyszukiwania potrzebnej informacji oraz wnioskowania na jej podstawie. Poszukiwane są nie tylko informacje zlokalizowane w pojedynczych dokumentach lecz również te, które są rozproszone po całym korpusie. Zapytania kierowane do tych korpusów mogą dotyczyć zarówno konkretnych faktów zawartych w tekstach, jak i ogólnych praw rządzących tymi faktami. Prawa te są formułowane w wyniku wnioskowania indukcyjnego (bazującego na generalizacji) i na ogół są prawdziwe jedynie z pewnym prawdopodobieństwem, którego wielkość trzeba oszacować.

W niniejszej pracy rozwijam metodologię realizującą powyższą funkcjonalność dla zbiorów tekstów o ograniczonej dziedzinie tematycznej. Założenie to pozwala reprezentować za pomocą ontologii strukturę informacji zawartej w tekstach. Opracowane metody testuję na dwu korpusach różniących się językiem i tematyką.

2 Istniejące rozwiązania dla danych z jawną strukturą

Problematyka wyszukiwania danych dotyczących zadanego zapytania została gruntownie zbadana w kontekście relacyjnych baz danych [9]. Zastosowane tam rozwiązania działają w oparciu o strukturę stanowiącą model danych, zwykle wyrażony za pomocą diagramu związków encji. Podobnie, zagadnienia generalizacji i wnioskowania aproksymacyjnego rozpatrywano w odniesieniu do danych reprezentowanych za pomocą tabeli zawierającej obiekty opisane za pomocą wektorów wartości atrybutów [8, 20, 34, 40].

Model danych pełni tu dwie role: wskazuje, gdzie znajdują się potrzebne dane oraz umożliwia wydobycie zawartej w nich informacji w celu dalszego przetwarzania.

3 Stan wiedzy dla problematyki związanej z danymi tekstowymi

Główną przeszkodą w analizie danych tekstowych jest brak danej wprost struktury informacji w nich zawartych. Rekonstrukcja jej jest konieczna, by odnaleźć dane w korpusie i wydobyć zawartą w nich informację. Aby wymodelować tę strukturę potrzebna jest wiedza o języku, w którym teksty są napisane i o świecie w nich opisanym. Przy czym, im dokładniej model pozwala wyrazić treść tekstu, tym więcej wiedzy trzeba do niego wprowadzić, co ogranicza zakres zastosowań wyrafinowanych modeli.

Przykładowo, wprowadzenie prostego modelu języka zwanego workiem słów [16] (stosowanego m.in. w wyszukiwarkach internetowych) nie wymaga wprowadzania wiedzy o treści tekstów i umożliwia wyszukanie dokumentów związanych z zapytaniem, ale nie pozwala na wyciąganie z nich wniosków ani na scalenie zawartych w nich informacji.

Ekstrakcja Informacji (*Information Extraction*) [35] realizuje proste zadania wydobywania ustrukturalizowanej informacji z tekstów, jednak zastosowana w niej metodologia obejmująca parsowanie powierzchniowe i nie umożliwiająca wyrażenia niejednoznaczności, nie jest skalowalna na przypadek bardziej skomplikowanej struktury.

Z kolei metody prowadzące do dokładnego oddania treści tekstu tworzone są w oparciu o analizę pojedynczych problemów semantycznych (np. reprezentacji kwantyfikowania, liczby mnogiej, czy mowy zależnej) [3, 7, 14, 36], bądź też niewielkich liczących poniżej 1000 zdań, zbiorów tekstów [1, 17, 45, 58].

Miały miejsce również próby sprowadzenia korpusów do postaci formuł logicznych, jedynie na podstawie wiedzy o języku, bez wykorzystania wiedzy o związkach pomiędzy pojęciami opisanymi w tekstach [4, 5, 11]. Uzyskane w ten sposób formuły logiczne odzwierciedlają treść tekstów, lecz nadal nie wyrażają w jawny sposób struktury informacji w nich zawartych. Na przykład nie pozwalają one utożsamić wypowiedzi zawierających tę samą treść a różniących się sposobem wyrażenia jej za pomocą języka.

4 Reprezentacja treści tekstów

Przedstawione powyżej wyniki wskazują na to, że konieczne jest stworzenie struktury reprezentującej informację zawartą w tekstach, a następnie dokonanie generalizacji ustrukturalizowanej informacji, tak by uzyskać jej podsumowanie zrozumiałe dla człowieka.

Metodologia zaproponowana w pracy opiera sie na założeniu, że **badane teksty dotyczą konkretnej dziedziny i mają ograniczony zakres tematyczny**. Dzięki temu możliwe jest formalne opisanie struktury informacji dotyczących tej dziedziny za pomocą ontologii. Ontologia [51] reprezentuje wiedzę dziedzinową: dzieli zbiór obiektów należących do badanej dziedziny na kategorie (typy) oraz określa związki pomiędzy tymi kategoriami. Ontologia jest modelem danych analogicznym do stosowanego w przypadku baz danych diagramu encji.

Podejście to, stosowane w sztucznej inteligencji i analizie danych [2, 18, 38, 49], nie było jeszcze wykorzystywane podczas przetwarzania języka. Niektóre systemy [4,

5, 11] korzystają z WordNetu [33] — bazy wiedzy o związkach pomiędzy znaczeniami słów, takimi jak synonimia, hiponimia, czy meronimia. Jednak uzyskana w ten sposób ontologia jest zbyt powierzchowna by wyznaczyć strukturę konkretnej dziedziny. Jej główną wadą jest to, że nie opisuje w jaki sposób pojęcia złożone są konstruowane z pojęć prostszych.

Informację wydobytą z dokumentów przechowuję jako zbiór formuł języka zwanego językiem reprezentacji znaczenia. Zdania w tym języku składają się z połączonych alternatywą i koniunkcją predykatów, których argumentami są stałe. Stałe reprezentują obiekty o których jest mowa w analizowanych tekstach. Predykaty reprezentują pojęcia z ontologii, wyrażają fakt przynależności obiektu do danej kategorii ontologicznej oraz opisane przez ontologię związki pomiędzy obiektami. Zakładam, że **teksty opisują jedynie zależności pomiędzy konkretnymi obiektami**. Dzięki temu mogę oznaczyć każdy obiekt za pomocą unikalnej stałej i wyrazić treść tekstów bez użycia zmiennych ani kwantyfikatorów.

Korzystam z pojęć teorii modeli, aby formalnie zdefiniować semantykę języka reprezentacji znaczenia. Informacja zawarta w dokumentach jest niekompletna i niejednoznaczna. Z tego względu semantykę dokumentu wyrażam jako klasę możliwych światów — modeli zgodnych z jego treścią. Teksty opisują własności realnych obiektów i to te obiekty konstytuują uniwersa możliwych światów. Oznacza to, że własności takie, jak przynależność do kategorii ontologicznej, są zdefiniowane w taki sam sposób we wszystkich możliwych światach.

Problem związku pomiędzy językiem i światem znany jest w filozofii języka jako problem odniesienia przedmiotowego. Przedstawione w pracy podejście jest podobne do obrazkowej teorii odniesienia przedmiotowego zaproponowanej przez Wittgensteina w [54].

5 Przetwarzanie tekstów

Zastosowany w pracy schemat przetwarzania języka jest wzorowany na powszechnie stosowanych metodach odwzorowywania znaczenia wypowiedzeń w języku naturalnym w formuły języka reprezentacji znaczenia [29]. Metody te dostosowałem do sytuacji, w której dostępna jest ontologia.

U podstaw zastosowanej w pracy metodologii leży założenie kompozycjonalności: zgodności struktury opisującej świat ze strukturą języka który go opisuje. Zakładam, że **składniowym operacjom tworzenia złożonych fraz na podstawie fraz prostszych odpowiada opisywanie złożonych obiektów za pomocą ich komponentów**.

Badania prowadzę w oparciu o dwa zbiory tekstów: korpus sumeryjskich tekstów gospodarczych z okresu III dynastii z Ur [6] oraz słownik biobibliograficzny współczesnych polskich pisarzy i badaczy literatury [12]. Dla każdego z korpusów opracowałem gramatykę, która opisuje w jaki sposób frazy są konstruowane ze słów i innych, prostszych fraz w poszczególnych zdaniach korpusu.

Z uwagi na to, że sumeryjskie teksty gospodarcze są niemal pozbawione informacji składniowych, przetwarzam je za pomocą gramatyki semantycznej, której symbole nieterminalne reprezentują kategorie ontologiczne bytów o których jest mowa w tekstach. Wytworzona dla tekstów sumeryjskich gramatyka pełni również rolę leksykonu:

reguły gramatyczne rozpoznają słowa z języka i przypisują im kategorie ontologiczne.

Z drugiej strony teksty polskie zawierają bogatą informację składniową. Dlatego w przypadku tych tekstów korzystam jednocześnie z dwu gramatyk: jedna analizuje cechy składniowe, a druga semantyczne. Rozwiązanie to, nieznane dotąd w literaturze przedmiotu, pozwoliło stworzyć prostą, uniwersalną gramatykę fragmentu języka polskiego. Wykorzystuje ona jedynie część informacji składniowych zawartych w tekstach, co powoduje, że jest istotnie prostsza niż gramatyki uwzględniające wszystkie niuanse językowe [37, 39, 46, 53, 56]. Z kolei gramatyka semantyczna, opisująca zależności pomiędzy kategoriami ontologicznymi obiektów rozpoznawanych przez poszczególne frazy, jest generowana bezpośrednio na podstawie wiedzy dziedzinowej zawartej w ontologii. Wymaganie jednoczesnej spójności składniowej i semantycznej istotnie redukuje niejednoznaczność wywodu gramatycznego. Współdziałanie obu gramatyk oparte jest na założeniu kompozycjonalności. Źródło wiedzy o fleksji języka polskiego stanowi analizator morfologiczny Morfeusz [55].

Korpus tekstów sumeryjskich przetwarzam za pomocą opracowanego w tym celu parsera, zaś w przypadku tekstów polskich korzystam z parsera tablicowego [15].

Podczas parsowania z każdym symbolem z gramatyki skojarzona jest wartość semantyczna: formuła języka reprezentacji znaczenia, opisująca semantykę fragmentu tekstu, na podstawie którego ten symbol został wygenerowany. Reguły gramatyczne są zaopatrzone w załączniki semantyczne (*semantic attachment*). Reguły gramatyki określają jakie symbole gramatyczne są generowane podczas parsowania, a załączniki semantyczne są funkcjami wyliczającymi wartości semantyczne generowanych symboli na podstawie semantyk symboli rozpoznanych przez regułę.

Dane tekstowe są niejednoznaczne. W szczególności uszkodzone sumeryjskie dokumenty gospodarcze można zinterpretować na olbrzymią ilość sposobów. W pracy nie próbuję redukować niejednoznaczności podczas parsowania z uwagi no to, że nadmiarowe interpretacje mogę znacznie precyzyjniej usunąć operując na formułach języka reprezentacji znaczenia. Alternatywne interpretacje łączę za pomocą alternatywy. Formuły semantyczne buduję zgodnie z procesem rozbioru składniowego, co pozwala mi w zwarty sposób reprezentować wszystkie możliwe interpretacje niejednoznacznego dokumentu, dzięki rozproszeniu ich pomiędzy symbolami gramatyki.

Reprezentacje semantyczne tekstów niejednoznacznych były badane przez teoretyków [10, 11, 47, 48], lecz nie stosowano ich w praktyce. W zastosowaniach praktycznych zakłada sie, że gramatyka jest jednoznaczna [57]. Wybiera się najbardziej prawdopodobne drzewo wywodu [4] lub przetwarza jedynie zdania, które mają tylko jedną interpretację [44, 45]. Związany z niejednoznacznością problem przetwarzania dokumentów uszkodzonych nie był jeszcze poruszany w literaturze przedmiotu.

Pokazuję również w jaki sposób można analizować teksty których fragmenty zawierają pojęcia nie należące do ontologii. Funkcjonalność tą uzyskuję domykając ontologię za pomocą specjalnej kategorii zawierającej wszystkie obiekty, które nie należą do pozostałych kategorii ontologicznych. Dzięki temu do przetwarzania tekstów wystarcza niewielka ontologia, nie zawierająca rzadkich, czy nieistotnych pojęć.

Uzyskaną w wyniku parsowania formułę języka reprezentacji znaczenia trzeba poddać dalszej analizie która uporządkuje niekompozycjonalne (w ujęciu zastosowanej gramatyki) fragmenty języka takie jak zdania względne, konstrukcje imiesłowowe, nominalizacje, czy wtrącenia; zastąpi spójniki języka naturalnego spójnikami logicz-

nymi oraz zredukuje niejednoznaczność. Redukcja niejednoznaczności opiera się na odrzuceniu interpretacji sprzecznych, wybraniu maksymalnej (pod względem zawartości informacyjnej) interpretacji oraz sprowadzeniu interpretacji do postaci kanonicznej. Należy tutaj zaznaczyć, że całkowite usunięcie niejednoznaczności nie jest możliwe i to właśnie jej obecność stanowi główną różnicę pomiędzy informacją zapisaną w postaci formuł języka reprezentacji znaczenia a relacyjną bazą danych.

Otrzymuję reprezentację treści dokumentów mającą jawną, wyznaczoną przez ontologię, strukturę. Dla tak przygotowanych danych konstruuję język zapytań. Każde zapytanie wyznacza pewien zbiór możliwych światów, a jako odpowiedź na nie wyszukiwane są dane, których semantyka jest podzbiorem zbioru wyznaczonego przez zapytanie. Na poziomie składni zapytania są wzorcami, a wyszukiwane są dane pasujące do wzorca.

6 Wnioskowanie indukcyjne

Zagadnienia generalizacji i wnioskowania aproksymacyjnego bada teoria zbiorów przybliżonych [41, 42]. W swoim standardowym ujęciu zakłada ona, że dane są informacje o obiektach reprezentowane w postaci tabeli wektorów wartości atrybutów tych obiektów. Okazuje się, że teorię zbiorów przybliżonych można elegancko rozszerzyć na przypadek danych zapisanych za pomocą formuł języka reprezentacji znaczenia.

Rozszerzenia dokonuję modyfikując definicje podstawowych pojęć teorii. Następnie dowodzę, że pierwotne definicje są szczególnym przypadkiem swoich rozszerzonych wersji. Pokazuję też zgodność zmodyfikowanej definicji z rozszerzeniami teorii na tabele zawierające niejednoznaczności wyrażone przez brakujące wartości oraz wielowartościowe atrybuty [13, 19, 30, 31, 32, 40]. Z punktu widzenia przetwarzania języka brakujące wartości odpowiadają niekompletności informacji zawartej w tekstach a wielowartościowe atrybuty są prostymi przypadkami niejednoznaczności.

Teoria zbiorów przybliżonych umożliwia stworzenie intuicyjnego opisu zależności występujących w dostępnych danych. Aby uzyskać opis prawdziwy dla obiektów, o których nie posiadamy informacji konieczne jest dodanie dodatkowych założeń o związku pomiędzy znanymi obiektami, a tymi, które będą obserwowane w przyszłości [43, 50].

Statystyczna teoria uczenia (*Statistical Learning Theory*) [52] wprowadza probabilistyczny model generowania danych. Wykorzystuję go, by oszacować jakość generalizacji i poprawność wyciągniętych wniosków za pomocą współczynników liczbowych.

Aby wyniki badań zastosować w praktyce opracowałem klasyfikator — algorytm znajdujący zależności pomiędzy wartościami cech obiektów. Klasyfikator ten wykorzystuje metody teorii zbiorów przybliżonych do modelowania zależności, a uzyskane wyniki ocenia korzystając z nierówności probabilistycznych [21, 22, 23] opartych na założeniach statystycznej teorii uczenia.

Zależności znajdowane przez klasyfikator są wyrażane za pomocą zbioru reguł. Każda reguła ma postać implikacji i jest zaopatrzona w oszacowanie prawdopodobieństwa prawdziwości następnika implikacji pod warunkiem prawdziwości jej poprzednika. Dzięki temu klasyfikator pozwala wyrażać w sposób zrozumiały dla człowieka globalne zależności zawarte w danych.

7 Testy

Opisaną w pracy metodologię przetestowałem na dwu korpusach liczących po około 800000 słów: korpusie sumeryjskich tekstów gospodarczych z okresu III dynastii z Ur [6] oraz słowniku biobibliograficznym współczesnych polskich pisarzy i badaczy literatury [12].

Eksperymenty wykazały, że wprowadzenie ontologii pozwala wypełnić lukę pomiędzy powierzchniowym a głębokim przetwarzaniem języka. W pracy przedstawiłem też wyniki szeregu eksperymentów opierających się na generalizacji ustrukturalizowanej informacji uzyskanej w wyniku przetwarzania języka. Zostały one pozytywnie ocenione przez ekspertów.

Część materiałów zawartych w pracy pochodzi z publikacji: [22, 23, 24, 25, 26, 27, 28].

Praca doktorska powstała przy wsparciu Ministerstwa Nauki i Szkolnictwa Wyższego z grantu promotorskiego N N206 400234 oraz z grantu N N516 368334.

Literatura

[1] ANDROUTSOPOULOS, I., RITCHIE, G. D., AND THANISCH, P. Natural Language Interfaces to Databases–an introduction. *Journal of Language Engineering 1*, 1 (1995), 29–81.

[2] BAZAN, J., NGUYEN, S. H., NGUYEN, H. S., AND SKOWRON, A. Rough set methods in approximation of hierarchical concepts. In *Proc. of the Fourth International Conference on Rough Sets and Current Trends in Computing (RSCTC 2004)* (Heidelberg, 2004), S. Tsumoto, R. Slowinski, J. Komorowski, and J. Grzymala-Busse, Eds., vol. 3066 of *Lecture Notes in Artificial Intelligence*, Springer-Verlag, pp. 346–355.

[3] BLACKBURN, P., AND BOS, J. *Representation and Inference for Natural Language: A First Course in Computational Semantics*. CSLI Publications, Stanford, California, 2005.

[4] BOS, J. Towards wide-coverage semantic interpretation. In *Proc. of Sixth International Workshop on Computational Semantics (IWCS-6)* (2005), pp. 42–53.

[5] BOS, J., CLARK, S., STEEDMAN, M., CURRAN, J. R., AND HOCKENMAIER, J. Wide-coverage semantic representations from a ccg parser. In *Proc. of the 20th International Conference on Computational Linguistics (COLING '04)* (Geneva, Switzerland, 2004), COLING, pp. 1240–1246.

[6] The cuneiform digital library initiative (CDLI), 2000–2008. `http://cdli.ucla.edu`.

[7] CHARNIAK, E., AND WILKS, Y., Eds. *Computational Semantics*. North-Holland / American Elsevier, Amsterdam, 1976.

[8] CICHOSZ, P. *Systemy uczące się [Machine learning systems]*. Wydawnictwa Naukowo–Techniczne, Warszawa, 2000.

[9] CODD, E. F. A relational model of data for large shared data banks. *Commun. ACM 13*, 6 (1970), 377–387.

[10] COPESTAKE, A., FLICKINGER, D., AND SAG, I. Minimal recursion semantics: an introduction. Tech. rep., CSLI, Stanford, CA, 1999.

[11] CROUCH, D. Packed rewriting for mapping semantics to KR. In *Proc. of Sixth International Workshop on Computational Semantics* (2005).

[12] CZACHOWSKA, J., AND SZAŁAGAN, A., Eds. *Współcześni polscy pisarze i badacze literatury. Słownik biobibliograficzny. [Contemporary Polish Writers and Literary Scholars. Biobibliographical Lexicon]*. Wyd. Szkolne i Pedagogiczne, last volumes Wyd. IBL PAN, 1994,1996,1997,1999,2001,2003,2004,2008.

[13] DEMRI, S., AND ORŁOWSKA, E. *Incomplete Information: Structure, Inference, Complexity*. Monographs in Theoretical Computer Science. An EATCS Series. Springer, 2002.

[14] DOWTY, D. R., WALL, R. E., AND PETERS, S. *Introduction to Montague Semantics*. Reidel, Dordrecht, 1981.

[15] EARLEY, J. An efficient context-free parsing algorithm. *Communications of the ACM 6*, 8 (1986), 451–455.

[16] FELDMAN, R., AND SANGER, J. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2006.

[17] GE, R., AND MOONEY, R. J. A statistical semantic parser that integrates syntax and semantics. In *Proc. of 9th Conf. on Computational Natural Language Learning (CoNLL-2005)* (Ann Arbor, MI, 2005), pp. 9–16.

[18] GRUBER, T. R. A translation approach to portable ontologies. *Knowledge Acquisition 5*, 2 (1993), 199–220.

[19] GRZYMAŁA-BUSSE, J. W., AND GRZYMAŁA-BUSSE, W. J. An experimental comparison of three rough set approaches to missing attribute values. *T. Rough Sets 6* (2007), 31–50.

[20] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. H. *The Elements of Statistical Learning*. Springer, 2001.

[21] HOEFFDING, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58* (1963), 13–30.

[22] JAWORSKI, W. Model selection and assessment for classification using validation. In *Proc. of Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 10th International Conference (RSFDGrC 2005)* (2005), D. Ślęzak, G. Wang, M. S. Szczuka, I. Düntsch, and Y. Yao, Eds., vol. 3641 of *Lecture Notes in Computer Science*, Springer, pp. 481–490.

[23] JAWORSKI, W. Bounds for validation. *Fundam. Inform 70*, 3 (2006), 261–275.

[24] JAWORSKI, W. Learning compound decision functions for sequential data in dialog with experts. In *Proc. of Rough Sets and Current Trends in Computing, 5th International Conference, (RSCTC 2006)* (2006), S. Greco, Y. Hata, S. Hirano, M. Inuiguchi, S. Miyamoto, H. S. Nguyen, and R. Slowinski, Eds., vol. 4259 of *Lecture Notes in Computer Science*, Springer, pp. 627–636.

[25] JAWORSKI, W. Automatic tool for semantic analysis of Neo-Sumerian documents. In *Proc. of 52nd Rencontre Assyriologique Internationale* (2007). (to appear).

[26] JAWORSKI, W. Contents modelling of Neo-Sumerian Ur III economic text corpus. In *Proc. of the 22nd International Conference on Computational Linguistics (COLING 2008)* (Manchester, UK, 2008), Coling 2008 Organizing Committee, pp. 369–376.

[27] JAWORSKI, W. Generalized indiscernibility relations: Applications for missing values and analysis of structural objects. *Transactions of Rough Sets 8* (2008), 116–145.

[28] JAWORSKI, W. Rule induction: Combining rough set and statistical approaches. In *Proc. of Rough Sets and Current Trends in Computing, 6th International Conference (RSCTC 2008)* (2008), C.-C. Chan, J. W. Grzymala-Busse, and W. Ziarko, Eds., vol. 5306 of *Lecture Notes in Computer Science*, Springer, pp. 170–180.

[29] JURAFSKY, D., AND MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey, 2000.

[30] KRYSZKIEWICZ, M. Rough set approach to incomplete information systems. *Inf. Sci 112*, 1-4 (1998), 39–49.

[31] LATKOWSKI, R. Flexible indiscernibility relations for missing attribute values. *Fundam. Inform 67*, 1-3 (2005), 131–147.

[32] LIPSKI, W. J. On Databases with Incomplete Information. *Journal of the Association of Computing Machinery 28*, 1 (1981), 41–70.

[33] MILLER, G. A., BECKWITH, R., FELLBAUM, C., GROSS, D., AND MILLER, K. Five papers on WORDNET. Tech. Rep. CSL 43, Cognitive Science Laboratory, Princeton University, 1990.

[34] MITCHELL, T. M. *Machine Learning*. McGraw-Hill, New York, 1997.

[35] MOENS, M.-F. *Information Extraction: Algorithms and Prospects in a Retrieval Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[36] MONTAGUE, R. Universal grammar. *Theoria 36* (1970), 373–398.

[37] MYKOWIECKA, A. *Inżynieria lingwistyczna. Komputerowe przetwarzanie tekstów w języku naturalnym [Linguistic Engineering. Computer processing of texts in natural language]*. Polsko-Japońska Wyższa Szkoła Technik Komputerowych, 2007.

[38] NGUYEN, S. H., BAZAN, J., SKOWRON, A., AND NGUYEN, H. S. Layered learning for concept synthesis. *LNCS Transactions on Rough Sets 3100*, 1 (2004), 187–208.

[39] OBRĘBSKI, T. *Automatyczna analiza składniowa języka polskiego z wykorzystaniem gramatyki zależnościowej [Automatic syntactic analysis of language using dependence grammar]*. PhD thesis, IPI PAN, Poznań, 2002.

[40] PAWLAK, Z. Information systems — theoretical foundations. *Information Systems 6*, 3 (1981), 205–218.

[41] PAWLAK, Z. Rough sets. *International Journal of Computer and Information Sciences 11*, 5 (1982), 341–356.

[42] PAWLAK, Z. *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht, 1991.

[43] PAWLAK, Z., AND SKOWRON, A. Rough sets: Some extensions. *Inf. Sci 177*, 1 (2007), 28–40.

[44] POPESCU, A.-M., ARMANASU, A., ETZIONI, O., KO, D., AND YATES, A. Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability. In *Proc. of the Twentieth International Conference on Computational Linguistics (COLING–04)* (2004), pp. 30–39.

[45] POPESCU, A.-M., ETZIONI, O., AND KAUTZ, H. Towards a theory of natural language interfaces to databases. In *Proc. of the 2003 International Conference on Intelligent User Interfaces* (2003), Full Technical Papers, pp. 149–157.

[46] PRZEPIÓRKOWSKI, A., KUPŚĆ, A., MARCINIAK, M., AND MYKOWIECKA, A. *Formalny opis języka polskiego. Teoria i implementacja [A formal description of the Polish language. Theory and implementation]*. Akademicka Oficyna Wydawnicza Exit, Warszawa, 2002.

[47] REYLE, U. Dealing with ambiguities by underspecification: Construction, representation, and deduction. *Journal of Semantics 10*, 2 (1993), 123–179.

[48] RICHTER, F., AND SAILER, M. Underspecified semantics in HPSG. In *Proc. of the Second International Workshop on Computational Semantics* (1999), Kluwer Academic Publishers, pp. 95–112.

[49] SKOWRON, A., AND STEPANIUK, J. Towards discovery of information granules. In *Proc. of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery (PKDD-99)* (Berlin, 1999), J. M. Zytkow and J. Rauch, Eds., vol. 1704 of *LNAI*, Springer, pp. 542–547.

[50] SKOWRON, A., ŚWINIARSKI, R., AND SYNAK, P. Approximation spaces and information granulation. *LNCS Transactions on Rough Sets 3400*, 3 (2005), 175–189.

[51] STAAB, S., AND STUDER, R., Eds. *Handbook on Ontologies*. Springer-Verlag, Heidelberg and Berlin, 2004.

[52] VAPNIK, V. N. *Statistical Learning Theory.* John_Wiley, 1998.

[53] VETULANI, Z. *Komunikacja człowieka z maszyną. Komputerowe modelowanie kompetencji językowej [Man-machine communication. Computer modeling of linguistic competence]*. Akademicka Oficyna Wydawnicza Exit, Warszawa, 2004.

[54] WITTGENSTEIN, W. *Tractatus Logico-Philosophicus*. Routledge and Kegan Paul, London, 1962.

[55] WOLIŃSKI, M. Morfeusz — a practical tool for the morphological analysis of polish. In *Proc. of Intelligent Information Processing and Web Mining (IIS:IIPWM'06)* (2006), M. Kłopotek, S. Wierzchoń, and K. Trojanowski, Eds., Springer, pp. 503–512.

[56] WOLIŃSKI, M. *Komputerowa weryfikacja gramatyki Świdzińskiego [Computer driven verification of Świdziński's grammar]*. PhD thesis, IPI PAN, Warszawa, 2004.

[57] WONG, Y., AND MOONEY, R. J. Learning for semantic parsing with statistical machine translation. In *Proc. of Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT–NAACL–06)* (New York City, NY, 2006), pp. 439–446.

[58] ZETTLEMOYER, L. S., AND COLLINS, M. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proc. of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)* (2005), AUAI Press, pp. 658–666.