Irodalomjegyzék

[1] L. Aas and L. Eikvil. Text categorisation: A survey. Raport NR 941, Norwegian Computing Center, 1999.

[2] Abonyi János, szerk. *Adatbányászat – a hatékonyság eszköze. Gyakorlati útmutató kezdőknek és haladóknak*. ComputerBooks, Budapest, 2005.

[3] Z. Alexin, J. Dombi, K. Fábricz, T. Gyimóthy, and T. Horváth. CONSTRUCTOR: A natural language interface based on attribute grammars. *Acta Cybernetica*, 9(3):247–255, 1990.

[4] J. Allen. *Natural Language Understanding*. The Benjaming/Cummings Publishing, Redwood City, USA, 2nd edition, 1995.

[5] H. Alshawi. *The Core Language Engine*. MIT Press, Cambridge, USA, 1992.

[6] H. Alshawi and J. van Eijck. Logical forms in the core language engine. In *Proc. of ACL-89, 27th Annual Meeting on Assoc. for Computational Linguistics*, pp. 25–32, Vancouver, Canada, 1989.

[7] I. Androutsopoulos, G. D. Ritchie, and P. Thanisch. Masque/SQL – an efficient and portable natural language query interface for relational databases. In *Proc. of IEA/AIE-93 Conf.*, pp. 327–330, Edinburgh, UK, 1993.

[8] I. Androutsopoulos, G. D. Ritchie, and P. Thanisch. Natural language interfaces to databases – an introduction. *J. of Natural Language Engineering*, 1(1):29–81, 1995.

[9] C. Apte, F. J. Damerau, and S. M. Weiss. Automated learning of decision rules for text categorization. *ACM Trans. Information Systems*, 12(3):233–251, 1994.

[10] S. E. Arnold. The Google Legacy, 2005.

[11] D. M. Ayuso. Discourse entities in janus. In *Proc. of ACL-89, 27th Annual Meeting on Assoc. for Computational Linguistics*, pp. 243–250, Vancouver, Canada, 1989.

[12] M. Bacchin, N. Ferro, and M. Melucci. University of Padua at CLEF 2002: Experiments to evaluate a statistical stemming algorithm. In *Working Notes for the CLEF 2002 Workshop*, Roma, Italy, 2002.

[13] Bach Iván. *Formális nyelvek*. TypoTEX, Budapest, 2. kiadás, 2002.

[14] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *Proc. of SIGIR-98, 21st ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 96–103, Melbourne, Australia, 1998.

[15] V. Balasubramanian. State of the art review on hypermedia issues and application. Report, E-Papyrus, 1993.

[16] B. W. Ballard, J. C. Lusth, and N. L. Tinkham. LDC-1: a transportable, knowledge-based natural language processor for office environments. *ACM Trans. on Information Systems*, 2(1):1–25, 1984.

[17] M. Bates, M. G. Moser, and D. Stallard. The IRUS transportable natural language database interface. In *Proc. of the 1st Int. Workshop on Expert Database Systems*, pp. 617–630, Kiawah Island, USA, 1984.

[18] F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In *Proc. of SIGKDD-02, 8th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 436–442, Edmonton, Canada, 2002.

[19] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proc. of SIGIR-98, 23rd ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 192–199, Athens, Greece, 2000.

[20] M. K. Bergman. The deep web: surfacing hidden value. *J. of Electronic Publishing*, 7(1), 2001.

[21] T. Berners-Lee. Information management: A proposal, 1989. CERN.

[22] M. W. Berry. *Survey of Text Mining*. Springer Verlag, New York, 2004.

[23] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1995.

[24] G. L. Berry-Rogghe and H. Wulz. An overview of PLIDIS, a problem solving information system with german as query language. In *Natural Language Communication with Computers*, number 63 in Lecture Notes in Computer Science, pp. 87–132. Springer, London, UK, 1978.

[25] A. Blum. Learning boolean functions in an infinite attribute space. *Machine Learning*, 9(4):373–386, 1992.

[26] R. Blumberg and S. Arte. The problem with unstructured data. *DM Review*, February 2003.

[27] L. Bolc, K. Kochut, A. Leśniewski, and T. Strzałkowski. Natural language information retrieval system dialog. In *Proc. of ACL-83, 1st Conf. on European Chapter of the Assoc. for Computational Linguistics*, pp. 196–203, Pisa, Italy, 1983.

[28] K. Bontcheva, M. Dimitrov, D. Maynard, V. Tablan, and H. Cunningham. Shallow methods for named entity coreference resolution. In *Proc. of TALN-02, Chaînes de références et résolveurs d'anaphores*, Nancy, France, 2002.

[29] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[30] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[31] P. D. Bruza and Th. P. van der Weide. Assessing the quality of hypertext views. *ACM SIGIR Forum*, 24(3):6–25, 1990.

[32] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[33] V. Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, 1945.

[34] W. B. Canvar and J. M. Trenkle. N-gram-based text categorization. In *Proc. of SDAIR-94, 3rd Annual Symp. on Document Analysis and Information Retrieval*, pp. 161–175, Las Vegas, USA, 1994.

[35] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. of SIGIR-98, 21 st ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 335–336, Melbourne, Australia, 1998.

[36] B. Carpenter. Phrasal queries with LingPipe and Lucene: ad hoc genomics text retrieval. *NIST Special Publication*, 2004.

[37] P. Cedermark. Swedish noun and adjective morphology in a natural language interface to databases. Master's thesis, Uppsala University, Department of Linguistics, 2003.

[38] S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, 7(3):163–178, 1998.

[39] N. A. Chinchor. MUC-7 named entity task definition. In *Proc. of MUC-7, 7th Message Understanding Conference*, 1998.

[40] W. W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. *Proc. of IJCAI-03, Int. Joint Conf. on Artificial Intelligence, Workshop on Information Integration on the Web*, pp. 73–78, 2003.

[41] W. W. Cohen and Y. Singer. Context sensitive learning methods for text categorization. *ACM Trans. Inform. Syst.*, 17(2):141–173, 1999.

[42] A. Copestake and K. Sparck-Jones. Natural language interfaces to databases. *Knowledge Engineering Review*, 5(4):225–249, 1990.

[43] F. Crestani and S. Wu. Testing the cluster hypothesis in distributed information retrieval. *Inf. Processing & Management*, 42(5):1137–1150, 2006.

[44] D. Crystal. *A nyelv enciklopédiája*. Osiris, Budapest, 2003.

[45] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of ACL-02, 40th Annual Meeting of the Assoc. for Computational Linguistics*, pp. 168–175, Philadelphia, USA, 2002.

[46] D. R. Cutting, J. O. Pedersen, D. Karger, and J. W. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proc. of SIGIR-92, 15th ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 318–329, Copenhagen, Denmark, 1992.

[47] Csendes Dóra, Hatvani Csaba, Alexin Zoltán, Csirik János, Gyimóthy Tibor, Prószéky Gábor és Váradi Tamás. Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. In *I. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY-03)*, pp. 238–245, Szeged, 2003.

[48] I. Dagan, Y. Karov, and D. Roth. Mistake-driven learning in text categorization. In *Proc. of EMNLP-97, 2nd Conf. on Empirical Methods in Natural Language Processing*, pp. 55–63, Providence, USA, 1997.

[49] S. D'Alessio, K. Murray, R. Schiaffino, and A. Kershenbaum. The effect of using hierarchical classifiers in text categorization. In *Proc. of RIAO-00, 6th Int. Conf. Recherche d'Information Assistee par Ordinateur*, pp. 302–313, Paris, France, 2000.

[50] J. L. Dawson. Suffix removal for word conflation. *Bulletin of the Assoc. for Literary & Linguistic Computing*, 2(3):33–46, 1974.

[51] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. of the Am. Soc. for Information Science*, 41(6):391–407, 1990.

[52] H. Deitel, P. Deitel, T. Nieto, T. Lin, and P. Sadhau. *XML: How to program*. Prentice-Hall, 2001.

[53] P. Domingos and M. J. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2–3):103–130, 1997.

[54] S. T. Dumais. Improving the retrieval information from external sources. *Behaviour Research Methods, Instruments and Computers*, 23(2):229–236, 1991.

[55] S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proc. of CIKM-98, 7th ACM Int. Conf. on Information and Knowledge Management*, pp. 148–155, Bethesda, USA, 1998.

[56] H. P. Edmundson. New methods in automatic extracting. *J. of the ACM*, 16(2):264–285, 1969.

[57] H. P. Edmundson and R. E. Wyllys. Automatic abstracting and indexing-survey and recommendations. *Communications of the ACM*, 4(5):226–234, 1961.

[58] J. Edwards, K. McCurley, and J. Tomlin. An adaptive model for optimizing performance of an incremental web crawler. In *Proc. of WWW-01, 10th Int. Conf. on World Wide Web*, pp. 106–113, Hong Kong, 2001.

[59] J. Eisenstein and R. Davis. Gesture improves coreference resolution. In *Proc. of HLT-NAACL, Human Language Technology Conf. of the NAACL*, pp. 37–40, New York, USA, 2006.

[60] B. Endres-Niggemeyer. Human-style WWW summarization. Technical report, University for Applied Sciences, Department of Information and Communication, 2000.

[61] K. Fábricz, Z. Alexin, T. Gyimóthy, and T. Horváth. THALES: a software package for plane geometry constructions with a natural language interface. In *Proc. of COLING-90, 13th Int. Conf. on Computational Linguistics*, pp. 44–46, Helsinki, Finland, 1990.

[62] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka. Automated categorization in the international patent classification. *ACM SIGIR Forum Archive*, 37(1):10–25, 2003.

[63] C. J. Fall, A. Törcsvári, P. Fievét, and G. Karetka. Additional readme information for WIPO-de autocategorization data set, March 2003.

[64] C. J. Fall, A. Törcsvári, and G. Karetka. Readme information for WIPO-alpha autocategorization training set, December 2002.

[65] W. Fan, L. Wallace, S. Rich, and Z. Zhang. Tapping the power of text mining. *Communications of the ACM*, 49(9):76–82, 2006.

[66] Farkas Richárd és Szarvas György. Nyelvfüggetlen tulajdonnév felismerő rendszer, és alkalmazása különböző domainekre. In *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY-03)*, pp. 22–31, Szeged, 2006.

[67] P. P. Filipe and N. J. Mamede. Databases and natural language interfaces. In *Proc. of JISBD-00, 5th Jornada de Ingeniería de Software y Bases de Dados*, pp. 321–332, Valladolid, Spain, 2000.

[68] W. B. Frakes and C. J. Fox. Strength and similarity of affix removal stemming algorithms. *ACM SIGIR Forum*, 37(1):26–30, 2003.

[69] A. Franz and B. Milch. Searching the web by voice. In *Proc. of COLING-02, 19th Int. Conf. on Computational Linguistics*, pp. 1213–1217, Taipei, Taiwan, 2002.

[70] Füstös László, Meszéna György és Simonné Mosolygó Nóra. *A sokváltozós adatelemzés statisztikai módszerei*. Akadémiai Kiadó, Budapest, 1986.

[71] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[72] B. Galitsky. *Natural Language Question Answering System*. Advanced Knowledge International, Adelaide, Australia, 2003.

[73] M. K. Ganapathiraju. Relevance of cluster size in MMR based summarizer. Technical Report 11-742, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA, 2002.

[74] J. M. Gawron, J. King, J. Lamping, E. Loebner, E. A. Paulson, G. K. Pullum, I. A. Sag, and T. Wasow. Processing english with a generalized phrase structure grammar. In *Proc. of ACL-82, 20th Conf. on Assoc. for Computational Linguistics*, pp. 74–81, Toronto, Canada, 1982.

[75] C. F. Goldfarb. *The SGML Handbook*. Oxford University Press, 1990.

[76] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: Sentence selection and evaluation metrics. In *Proc. of SIGIR-99, 22nd ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 121–128, Berkeley, USA, 1999.

[77] J. Goldstein, V. Mittal, and J. Carbonell. Creating and evaluating multi-document sentence extract summaries. In *Proc. of CIKM-00, 9th Int. Conf. on Information Knowledge Management*, pp. 165–172, McLean, USA, 2000.

[78] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proc. of SIGIR-01, 24th ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 19–25, New Orleans, USA, 2001.

[79] G. Grefenstette and P. Tapanainen. What is a word, what is a sentence? Problems of tokenization. In *Proc. of COMPLEX-94, 3rd Int. Conf. on Computational Lexicography*, pp. 79–87, Budapest, Hungary, 1994.

[80] R. Grisham. Information extraction. In *The Oxford Handbook of Computational Linguistics*, pp. 545–559. Oxford University Press, 2004.

[81] F. Halasz and M. Schwartz. The Dexter hypertext reference model. *Communications of the ACM*, 37(2):30–39, 1994.

[82] P. Halácsy. Benefits of deep NLP-based lemmatization for information retrieval. In *Working Notes for the CLEF 2006 Workshop*, Alicante, Spain, 2006.

[83] P. Halácsy, A. Kornai, L. Németh, A. Rung, I. Szakadát, and V. Trón. Creating open language resources for Hungarian. In *Proc. of LREC-04, 4th Int. Conf. on Language Resources and Evaluation*, pp. 203–210, Lisbon, Portugal, 2004.

[84] E.-H. Han, D. Boley, M. Gini, R. Gross, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore. WebAce: A web agent for document categorization and exploration. In *Proc. of Agents-98, 2nd Int. Conf. on Autonomous Agents*, pp. 408–415, Minneapolis, USA, 1998.

[85] G. G. Hendrix, E. D. Sacerdoti, D. Sagalowicz, and J. Slocum. Developing a natural language interface to complex data. *ACM Trans. on Database Systems*, 3(2):105–147, 1978.

[86] R. L. Herschman, R. T. Kelly, and H. G. Miller. User performance with a natural language query system for command control. Technical Report NPRDC-TR-797, Navy Personnel Research and Development Center, San Diego, USA, 1979.

[87] W. Hersh, C. Buckley, T. Leone, and D. Hickman. OHSUMED: an interactive retrieval evaluation and new large text collection for research. In *Proc. of SIGIR-94, 17th ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 192–201, Dublin, Ireland, 1994.

[88] W. Howe. A brief history of the internet, 2005.

[89] D. A. Hull. Improving text retrieval for the routing problem using latent semantic indexing. In *Proc. of SIGIR-94, 17th ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 282–289, Dublin, Ireland, 1994.

[90] H. Jing and K. R. McKeown. The decomposition of human-written summary sentences. In *Proc. of SIGIR-99, 22nd ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 129–136, Berkeley, USA, 1999.

[91] T. Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proc. of ICML-97, 14th Int. Conf. on Machine Learning*, pp. 143–151, Nashville, USA, 1997.

[92] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. Technical Report, University of Dortmund, Dept. of Informatics, Dortmund, Germany, 1997.

[93] H.-M. Jung and G. G.-B. Lee. Multilingual question answering with high portability on relational databases. In *Proc. of COLING-02, Workshop on Multilingual Summarization and Question Answering*, pp. 1–8, Taipei, Taiwan, 2002.

[94] Zs. T. Kardkovács, D. Tikk, and Z. Bánsághi. The Ferrety algorithm for the KDD cup 2005 problem. *ACM SIGKDD Explorations Newsletter*, 7(2):111–116, 2005.

[95] B. Katz. *Using English for Indexing and Retrieving*, volume 1 of *Artificial Intelligence at MIT: Expanding Frontiers*, pp. 134–165. MIT Press, 1990.

[96] B. Katz, S. Felshin, D. Yuret, A. Ibrahim, J. J. Lin, G. Marton, A. J. McFarland, and B. Temelkuran. Omnibase: A uniform access to heterogeneous data for question

answering. In *Proc. of NLDB-02*, volume 2553 of *Lecture Notes in Computer Science*, pp. 230–234. Springer, Stockholm, Sweden, 2002.

[97] B. Katz and J.L. Lin. Start and beyond. In *Proc. of SCI 2002*, volume XVI of *World Multiconference on Systemics, Cybernetics and Informatics*, 2002.

[98] J. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. Introduction to the bio-entity recognition task at JNLPBA. In *Proc. of JNLPBA, Int. Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 70–75, Geneva, Switzerland, 2004.

[99] D. Koller and M. Sahami. Hierarchically classifying documents using a very few words. In *Proc. of ICML-97, 14th Int. Conf. on Machine Learning*, pp. 170–178, Nashville, USA, 1997.

[100] Korcsmáros István. Szövegbányászat (text mining) — új fogalom az üzleti intelligencia témakörében, 2003.

[101] Kornai András, Rebrus Péter, Vajda Péter, Halácsy Péter, Rung András és Trón Viktor. Általános célú morfológiai elemző kimeneti formalizmusa. In *II. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY-04)*, pp. 172–176, Szeged, 2004.

[102] M. Koskela, J. Laaksonen, and E. Oja. Entropy-based measures for clustering and SOM topology preservation applied to content-based image indexing and retrieval. In *Proc. of ICPR-04, 17th Int. Conf. on Pattern Recognition*, volume 2, pp. 1005–1008, 2004.

[103] Kovács László. *Adatbázisok tervezésének és kezelésének módszertana*. Computerbooks, 2004.

[104] M. Krier and F. Zaccà. Automatic categorization applications at the European Patent Office. *World Patent Information*, 24:187–196, 2002.

[105] R. Krovetz. Viewing morphology as an inference process. In *Proc. of SIGIR-93, 16th ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 191–203, Pittsburgh, USA, 1993.

[106] D. Küpper, M. Strobel, and D. Rösner. NAUDA: A cooperative natural language interface to relational databases. *ACM SIGMOD Records*, 22(2):529–533, 1993.

[107] D. Kuropka. *Modelle zur Repräsentation natürlichsprachlicher Dokumente. Ontologie-basiertes Information-Filtering und -Retrieval mit relationalen Datenbanken*. Logos Verlag, Berlin, 2004.

[108] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML-01, 18th Int. Conf. on Machine Learning*, pp. 282–289, Williamstown, USA, 2001.

[109] W. Lam and C. Y. Ho. Using a generalized instance set for automatic text categorization. In *Proc. of SIGIR-98, 21st ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 81–89, Melbourne, Australia, 1998.

[110] T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[111] K. Lang. Newsweeder: learning to filter netnews. In *Proc. of ICML-95, 12th Int. Conf. on Machine Learning*, pp. 331–339, Lake Tahoe, USA, 1995.

[112] L. S. Larkey. A patent search and classification system. In *Proc. of DL-99, 4th ACM Conf. on Digital Libraries*, pp. 179–187, Berkeley, USA, 1999.

[113] H.-D. Lee and J. C. Park. Interpretation of natural language queries for relational database access with combinatory categorial grammar. *Int. J. of Computer Processing of Oriental Languages*, 15(3):281–303, 2002.

[114] Lejtovicz Katalin és Kardkovács Zsolt T. Anaforafeloldás magyar nyelvű szövegekben. In *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY-06)*, pp. 362–363, Szeged, 2006.

[115] D. D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *Proc. of SIGIR-92, 15th ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 37–50, Copenhagen, Denmark, 1992.

[116] D. D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proc. of ECML-98, 10th European Conference on Machine Learning*, pp. 4–15, Chemnitz, Germany, 1998.

[117] D. D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *J. of Machine Learning Research*, 5:361–397, 2004.

[118] Y. H. Li and A. K. Jain. Classification of text documents. *The Computer Journal*, 41(8):537–546, 1998.

[119] R. Liere and P. Tadepalli. Active learning with committees for text categorization. In *Proc. of AAAI-97, 14th Conf. of the Am. Assoc. for Artificial Intelligence*, pp. 591–596, Providence, USA, 1997.

[120] N. Littlestone. Learning quickly when irrelevant attributes around: A new linear-threshold algorithm. *Machine Learning*, pp. 285–318, 1988.

[121] N. Littlestone. Comparing several linear-threshold learning algorithm on tasks involving superfluous attributes. In *Proc. of ICML-95, 12th Int. Conf. on Machine Learning*, pp. 353–361, Tahoe City, USA, 1995.

[122] P. C. Lockemann and F. B. Thompson. REL: a rapidly extensible language system. In *Proc. of Conf. on Computational Linguistics*, pp. 1–32, Sång-Säby, Sweden, 1969.

[123] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computation Linguistics*, 11(1):23–31, 1968.

[124] I. Mani. *Automatic Summarization*. John Benjamin's Publishing Company, 2001.

[125] Ch. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2007.

[126] Ch. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, 1999.

[127] A. McCallum. MALLET: A machine learning for language toolkit, 2002.

[128] A. Mccallum. Information extraction: Distilling structured data from unstructured text. *ACM Queue*, 3(9), 2005.

[129] A. McCallum, D. Freitag, and F. C. N. Pereira. Maximum entropy Markov models for information extraction and segmentation. In *Proc. of ICML-00, 17th Int. Conf. on Machine Learning*, pp. 591–598, Stanford, USA, 2000.

[130] A. McCallum, R. Rosenfeld, T. Mitchell, and A. Ng. Improving text classification by shrinkage in a hierarchy of classes. In *Proc. of ICML-98, 15th Int. Conf. on Machine Learning*, pp. 359–367, Madison, USA, 1998.

[131] C. T. Meadow, B. R. Boyce, and D. H. Kraft. *Text Information Retrieval Systems*. Academic Press, Orlando, 2nd edition, 2000.

[132] X.-F. Meng and S. Wang. Nchiql: The chinese natural language interface to databases. In *Proc. of DEXA-01, 12th Int. Conf. on Database and Expert Systems Applications*, volume 2113 of *Lecture Notes in Computer Science*, pp. 145–154. Springer, London, UK, 2001.

[133] T. M. Mitchell. *Machine Learning*. McGraw Hill, New York, 1996.

[134] R. Mitkov, editor. *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Gosport, Hampshire, 2004.

[135] J. L. Neto, A. D. Santos, C. A. A. Kaestner, and A. A. Freitas. Document clustering and text summarization. In *Proc of PADD-00, 4th Int. Conf. on Practical Applications of Knowledge Discovery and Data Mining*, pp. 41–55, London, UK, 2000.

[136] Németh László. Huntoken 1.4 kézikönyv, 2003.

[137] W. C. Ogden and P. Bernick. Using natural language interfaces. Technical Report MCCS-96-229, New Mexico State University, 1996.

[138] S. Olsen. IBM sets out to make sense of the Web, 2004.

[139] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[140] C. D. Paice. Another stemmer. *SIGIR Forum*, 24:56–61, 1990.

[141] L. A. F. Park. *Spectral Based Information Retrieval*. PhD thesis, The University of Melbourne, 2003.

[142] W. H. Paxton. *A Framework for Speech Understanding*. PhD thesis, SRI Artificial Intelligence Center, Menlo Park, USA, 1977.

[143] Ch. R. Perrault and B. J. Grosz. Natural-language interfaces. *Annual Review of Computer Science*, 1:47–82, 1986.

[144] J. L. Peterson. *Petri Net Theory and the Modelling of Systems*. Prentice-Hall, 1981.

[145] B. Pinkerton. Finding what people want: Experiences with the WebCrawler. In *Proc. of 2nd Int. WWW Conf.*, pp. 708–716, 1994.

[146] Pléh Csaba. Mondatmegértés a magyar nyelvben. In *Mondatközi viszonyok feldolgozása: az anafora megértése a magyarban*, pp. 164–195. Osiris, 2004.

[147] S. P. Ponzetto and M. Strube. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proc. of HLT-NAACL, Human Language Technology Conf. of the NAACL*, pp. 192–199, New York, USA, 2006.

[148] A.-M. Popescu, A. Armanasu, O. Etzioni, D. Ko, and A. Yates. Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability. In *Proc. of COLING-04, 20st Int. Conf. on Computational Linguistics*, pp. 141–147, 2004.

[149] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[150] M. F. Porter. Snowball: A language for stemming algorithms, 2001.

[151] Prószéky Gábor. *Számítógépes nyelvészet*. Számalk, Budapest, 1989.

[152] Prószéky Gábor. NewsPro: automatikus információszerzés gazdasági rövidhírekből. In *I. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY-03)*, pp. 161–167, Szeged, 2003.

[153] D. Radev, S. Blair-Goldensohn, and Z. Zhang. Experiments in single and multidocument summarization using MEAD. In *Proc. of DUC-01, Document Understanding Conf., Workshop on Text Summarization*, New Orleans, USA, 2001.

[154] R. A. P. Rangel, A. F. Gelbukh, J. J. G. Barbosa, E. A. Ruiz, A. M. Mejía, and A. P. D. Sánchez. Spanish natural language interface for a relational database querying system. In *Proc. of TSD-02, 5th Conf. on Text, Speech, and Dialogue*, volume 2448 of *Lecture Notes in Computer Science*, pp. 123–130. Springer, Brno, Czech Republic, 2002.

[155] P. Reis, J. Matias, and N. Mamede. Edite: A natural language interface to databases – A new perspective for an old approach. In *Proc. of ENTER-97, Information and Communication Technologies in Tourism*, pp. 317–326, Edinburgh, UK, 1997.

[156] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System — Experiments in Automatic Document Processing*, pp. 313–323. Prentice-Hall, Englewood Cliffs, 1971.

[157] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in brain. *Psychological Review*, pp. 386–407, 1958. (Újranyomva: *Neurocomputing*, MIT Press, 1988).

[158] D. Roth. Learning to resolve natural language ambiguities: a unified approach. In *Proc. of AAAI-98, 15th Conf. of the Am. Assoc. for Artificial Intelligence*, pp. 806–813, Madison, USA, 1998.

[159] J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, Ch. R. Johnson, and J. Scheffczyk. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, 2006.

[160] G. Salton, editor. *The SMART Retrieval System — Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, 1971.

[161] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1998.

[162] G. Salton and M. J. McGill. *An Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[163] E. F. T K. Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proc. of CoNLL-03, 7th Conf. on Computational Natural Language Learning*, pp. 142–147, Edmonton, Canada, 2003.

[164] R. E. Schapire and Y. Singer. BoosTexter: a boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[165] R. E. Schapire, Y. Singer, and A. Singhal. Boosting and Rocchio applied to text filtering. In *Proc. of SIGIR-98, 21st ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 215–223, Melbourne, Australia, 1998.

[166] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

[167] P. Schönhofen and H. Charaf. Sentence-based document size reduction. In *Proc. of ClusWEB-04, Clustering Information over the Web*, pp. 77–86, Heraklion, Greece, 2004.

[168] H. Schütze, D. A. Hull, and J. O. Pedersen. A comparison of classifiers and document representations for the routing problem. In *Proc. of SIGIR-95, 18th ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 229–237, Seattle, USA, 1995.

[169] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[170] F. Sebastiani, A. Sperduti, and N. Valdambrini. An improved boosting algorithm and its application to automated text categorization. In *Proc. of CIKM-00, 9th ACM Int. Conf. on Information and Knowledge Management*, pp. 78–85, McLean, USA, 2000.

[171] S. Sekine, K. Sudo, and C. Nobata. Extended named entity hierarchy. In *Proc. of LREC-02, 3rd Int. Conf. on Language Resources and Evaluation*, pp. 1818–1824, 2002.

[172] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang. Q^{2}C@UST: Our winning solution to query classification in KDD cup 2005. *ACM SIGKDD Explorations Newsletter*, 7(2):100–110, 2005.

[173] T. Sibanda and Ö. Uzuner. Role of local context in automatic deidentification of ungrammatical, fragmented text. In *Proc. of HLT-NAACL, Human Language Technology Conf. of the NAACL*, pp. 65–73, New York, USA, 2006.

[174] C. Silverstein, H. R. Henzinger, H. Marais, and M. Moricz. Analysis of a very large web search engine query log. *ACM SIGIR Forum*, 33(1):6–12, 1999.

[175] J. Slocum. A practical comparison of parsing strategies. In *Proc. of ACL-81, 19th Annual Meeting Assoc. for Computational Linguistics*, pp. 1–6, Stanford, USA, 1981.

[176] L. I. Smith. A tutorial on principal component analysis. Technical report, University of Otago, New Zealand, 2002.

[177] W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[178] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36(6):779–808, 2000.

[179] M. Steinbach, Karypis G., and V. Kumar. A comparison of document clustering techniques. In *Proc. of Text Mining Workshop at SIGKDD-00, 6th Int. Conf. on Knowledge Discovery and Data Mining*, Boston, USA, 2000.

[180] P. D. Stotts and R. Furuta. Petri-net-based hypertext: Document structure with browsing semantics. *ACM Trans. on Information Systems*, 7(1):3–29, 1989.

[181] Gy. Szarvas, R. Farkas, L. Felföldi, A. Kocsor, and J. Csirik. A highly accurate named entity corpus for Hungarian. In *Proc. of LREC-06, 5th Int. Conf. on Language Resources and Evaluation*, Genoa, Italy, 2006.

[182] M. Templeton and J. Burger. Problems in natural-language interface to DBMS with examples from EUFID. In *Proc. of the 1st Conference on Applied Natural Language Processing*, pp. 3–16, Santa Monica, USA, 1983.

[183] Tikk Domonkos, Biró György, Szidarovszky Ferenc P., Kardkovács Zsolt T., Héder Mihály és Lemák Gábor. Magyar internetes gazdasági tematikájú tartalmak keresése. In *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY-06)*, pp. 3–14, Szeged, 2006.

[184] D. Tikk, Gy. Biró, and A. Törcsvári. A hierarchical online classifier for patent categorization. In H. A. do Prado and E. Ferneda, editors, *Emerging Technologies of Text Mining: Techniques and Applications*. Idea Group Inc., 2007. (nyomdában).

[185] D. Tikk and Gy. Biró. Experiments with multilabel text classifier on the Reuters collection. In *Proc. of ICCC-03, 1st Int. Conf. on Computational Cybernetics*, pp. 33–38, Siófok, Hungary, 2003.

[186] D. Tikk, Gy. Biró, and J. D. Yang. Experiments with a hierarchical text categorization method on WIPO patent collections. In N. O. Attok-Okine and B. M. Ayyub, editors, *Applied Research in Uncertainty Modelling and Analysis*, number 20 in Int. Series in Intelligent Technologies, pp. 283–302. Springer, 2005.

[187] D. Tikk, Zs. T. Kardkovács, Z. Andriska, G. Magyar, A. Babarczy, and I. Szakadát. Natural language question processing for Hungarian deep web searcher. In *Proc. of ICCC-04, 2nd IEEE Int. Conf. on Computational Cybernetics*, pp. 303–309, Vienna, Austria, 2004.

[188] D. Tikk, Zs. T. Kardkovács, G. Magyar, A. Babarczy, and I. Szakadát. Natural language module of a Hungarian deep web searcher. In *Proc. of ISDA-04, 4th IEEE Int. Conf. on Intelligent Systems Design and Applications*, pp. 73–77, Budapest, Hungary, 2004.

[189] D. Tikk, Zs. T. Kardkovács, and F. P. Szidarovszky. Voting with a parameterized veto strategy: Solving the KDD cup 2006 problem by means of a classifier committee. *ACM SIGKDD Explorations Newsletter*, 8(2):53–62, 2006.

[190] Tikk Domonkos, Törcsvári Attila, Biró György és Bánsághi Zoltán. Szótövező eljárások hatása magyar szövegek automatikus kategorizálásánál. In *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY-05)*, pp. 430–434, Szeged, 2005.

[191] S. Tomlinson. Finnish, Portugese and Russian retrieval with Hummingbird SearchServer at CLEF 2004. In *Working Notes for the CLEF 2004 Workshop*, Bath, UK, 2004.

[192] A. Tordai and M. de Rijke. Hungarian monolingual retrieval at CLEF 2005. In *Working Notes for the CLEF 2005 Workshop*, Vienna, Austria, 2005.

[193] Trón Viktor. HunLex – morfológiai szótárkezelő rendszer. In *II. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY-04)*, pp. 177–182, Szeged, 2004.

[194] Trón Viktor, Halácsy Péter, Rebrus Péter, Rung András, Simon Eszter és Vajda Péter. Morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In *III. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY-05)*, pp. 169–179, Szeged, 2005.

[195] V. Trón, L. Németh, P. Halácsy, A. Kornai, Gy. Gyepesi, and D. Varga. Hunmorph: open source word analysis. In *Proc. of ACL Workshop on Software at the 43rd Annual Meeting of ACL*, 2005.

[196] K. Tumer and J. Ghosh. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3–4):385–403, 1996.

[197] L. Underwood. A brief history of search engines, 2005.

[198] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.

[199] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[200] Varasdi Károly. Koreferenciák feloldása, 2006. Magyar WordNet Projekt, projekt beszámoló.

[201] Vivisimo.com. How the Vivísimo clustering engine works, 2002.

[202] D. L. Waltz. An english language question answering system for a large relational database. *Communications of the ACM*, 21(7):526–539, 1978.

[203] D. H. D. Warren and F. C. N. Pereira. An efficient easily adaptable system for interpreting natural language queries. *Computational Linguistics*, 8(3–4):110–122, 1982.

[204] R. M. Weischedel, R. J. Bobrow, D. Ayuso, and L. Ramshaw. Portability in the janus natural language interface. In *Proc. of HLT-89, Workshop on Speech and Natural Language*, pp. 112–117, Philadelphia, USA, 1989.

[205] R. M. Weischedel, E. Walker, D. Ayuso, J. de Bruin, K. Koile, L. Ramshaw, and V. Shaked. Out of the laboratory: a case study with the IRUS natural language interface. In *Proc. of HLT-86, Workshop on Strategic Computing Natural Language*, pp. 44–61, Marina del Rey, USA, 1986.

[206] S. M. Weiss, C. Apte, F. J. Damerau, D. E. Johnson, F. J. Oles, T. Goetz, and T. Hampp. Maximizing text-mining performance. *IEEE Intelligent Systems*, 14(4):2–8, 1999.

[207] S. M. Weiss, N. Indurkhyra, T. Zhang, and F. J. Damerau. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer, 2004.

[208] W. Wibovo and H. E. Williams. Simple and accurate feature selection for hierarchical categorisation. In *Proc. of DE-02, ACM Symposium on Document Engineering*, pp. 111–118, McLean, USA, 2002.

[209] E. D. Wiener, J. O. Pedersen, and A. S. Weigend. A neural network approach to topic spotting. In *Proc. of the SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, pp. 317–332, Las Vegas, NV, 1995.

[210] W. Winiwarter. *The Integrated Deductive Approach to Natural Language Interfaces*. PhD thesis, Department of Information Engineering, University of Vienna, Austria, 1994.

[211] M. J. Witbrock and V. Mittal. Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries. In *Proc. of SIGIR-99, 22 nd ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 315–316, Berkeley, USA, 1999.

[212] M. B. Wolfe, M. E. Schreiner, B. Rehder, D. Laham, P. W. Foltz, W. Kintsch, and T. K. Landauer. Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes*, 25(2/3):309–336, 1998.

[213] W. A. Woods. Context-sensitive parsing. *Communications of the ACM*, 13(7):437–445, 1970.

[214] W. A. Woods, R. M. Kaplan, and B. L. Nash-Webber. The lunar sciences natural language information system. Technical Report BBN Report No. 2378, Bolt, Beranek, and Newman Inc., Cambridge, USA, 1972.

[215] Y. Yang. Expert network: effective and efficient learning from human decisions in text categorisation and retrieval. In *Proc. of SIGIR-94, 17th ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 13–22, Dublin, Ireland, 1994.

[216] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1–2):69–90, 1999.

[217] Y. Yang and C. G. Chute. An example-based mapping method for text categorization and retrieval. *ACM Trans. Inform. Syst.*, 12(3):252–277, 1994.

[218] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proc. of SIGIR-99, 22nd ACM Int. Conf. on Research and Development in Information Retrieval*, pp. 42–49, Berkeley, USA, 1999.

[219] Y. Yang and J. P. Pedersen. Feature selection in statistical learning of text categorization. In *Proc. of ICML-97, 14th Int. Conf. on Machine Learning*, pp. 412–420, Nashville, USA, 1997.

[220] O. Zamir, O. Etzioni, O. Madani, and R. M. Karp. Fast and intuitive clustering of web documents. In *Proc. of SIGKDD-97, 3rd Int. Conf. on Knowledge Discovery and Data Mining*, pp. 287–290, Newport Beach, USA, 1997.

[221] T. Zhang and F. J. Oles. Text categorization based on regularized linear classification methods. *Information Retrieval*, 4(1):5–31, 2001.