Extragere de cunoştinţe din texte în limba română şi date structurate cu aplicaţii în domeniul medical

Maria Mitrofan (Carp)

Conducători științifici:

acad. Ioan Dan Tufiș

acad. Constantin Ionescu-Târgovişte

2019

Cuprins

Introducere

Bazele extragerii de informații din texte (eng. Information Extraction, IE) au fost puse începând cu 1996, în cadrul conferințelor Message Understanding Conferences (MUCs) [1]. Extragerea informațiilor din texte constă în identificarea automată a informațiilor specifice legate de un subiect selectat dintr-un corpus. Prin identificarea entităților denumite, a evenimentelor și a relațiilor dintre ele s-a reușit extragerea informațiilor din diverse domenii (de exemplu terorismul din America Latină, pentru a identifica modelele legate de activitățile teroriste (MUC-4) [1]). O altă utilizare a tehnologiilor de IE este extragerea cunoștințelor sau a informațiilor din texte nestructurate. Astfel, extragerea informațiilor devine importantă pentru a face mai ușor accesul la fișierele de acest tip.

În domeniul biomedical apariția unor volume mari de date a accelerat în mod semnificativ cercetarea asupra domeniului. Cum o mare parte din datele disponibile în acest domeniu se găsesc într-o formă nestructurată, tehnicile de extragere a informațiilor din texte sunt utilizate pentru extragerea eficientă și automată a datelor și a relațiilor semnificative. Pentru a aborda această problemă au fost făcute studii riguroase de aplicare a IE la datele biomedicale. Astfel de eforturi de cercetare au început să poarte denumirea de mineritul literaturii biomedicale ([2, 3]). Deși de-a lungul timpului au fost dezvoltate o serie de intrumente și resurse utilizate în extragerea informațiilor din texte, în special pentru limba engleză, în foarte multe cazuri acestea nu sunt portabile în funcție de domeniu sau de limbă. În cazul limbii române, în domeniul biomedical nu au fost identificate resursele necesare (corpusuri însoțite de diverse

[1] https://cs.nyu.edu/faculty/grishman/muc6.html - accesat la 19.06.2018

tipuri de adnotări) antrenării sistemelor de extragere a informațiilor din textele medicale. Prin urmare și cercetările în acest domeniu sunt restrânse. Unul dintre principalele obiective ale stagiului doctoral este crearea resurselor necesare în extragerea de informații din textele medicale, fără de care cercetările în acest domeniu sunt dificile sau chiar imposibile (vezi capitolele 6 și 8).

1.1 Obiective

Principalele obiective ale tezei sunt:

1. Cercetarea rezultatelor actuale și a celor mai relevante aplicații în domeniu, dar și a standardelor de creare a resurselor specifice domeniului biomedical.
2. O1: Crearea unui corpus biomedical al limbii române.
3. O2: Adoptarea unui standard de adnotare cu entități denumite și crearea unei proceduri de adnotare.
4. O3: Crearea unui corpus biomedical „gold standard” adnotat la nivel morfologic și cu entități denumite.
5. O4: Adaptarea sistemelor de procesare a limbajului natural la domeniul biomedical.

1.2 Metodologie

În continuare este prezentată metodologia utilizată în cercetarea doctorală:

1. A fost studiată literatura de specialitate (cărți, articole științifice, pagini web) pentru o mai bună înțelegere a domeniului și a direcțiilor de cercetare existente, accentul fiind pus pe limba română.

2. Corpusul biomedical al limbii române (BioRo) a fost construit urmându-se procedura adoptată în cadrul proiectului CoRoLa [4].

3. Pentru reprezentarea entităților denumite a fost ales standardul IOB, acesta fiind cel mai utilizat la ora actuală în marcarea entităților denumite în textele biomedicale. Rețeaua semantică Unified Medical Language System (UMLS) a fost utilizată pentru a stabili grupurile și tipurile semantice de entități denumite utilizate în adnotarea corpusului.

4. Crearea corpusului MoNERo (corpus medical "Gold Standard" în limba română adnotat la nivel morfo-sintactic și cu entități denumite). Corpusul a fost adnotat atât morfologic automat, iar apoi corectat manual utilizându-se un set de 714 etichete cât și cu patru tipuri de entități denumite specifice domeniului medical.

5. Adaptarea sistemelor de adnotare la domeniul biomedical s-a făcut pe baza resurselor create, care au fost utilizate în antrenare și testare. Două tipuri de abordări bazate pe rețele neuronale au fost testate pentru adaptarea acestora la domeniul biomedical.

1.3 Prezentarea tezei

Această teză de doctorat este structurată în 7 capitole, excluzând introducerea și concluziile finale. Capitolele 2-7 prezintă documentarea teoretică premergătoare necesară în atingerea obiectivelor propuse și prezentate în capitolul de contribuții. Fiecare dintre aceste capitole teoretice evidențiază atât cadrul teoretic cât și metodologia de lucru necesare în dezvoltarea de resurse specifice prelucrării limbajului natural.

Teza conține 14 tabele, 11 figuri, un glosar de termeni și aproximativ 200 de referințe. Exemplele prezentate în teză sunt selectate cu precădere din corpusul BioRo.

Capitolul 2 prezintă principalele noțiuni teoretice necesare în procesul de construire a unui corpus. În secțiunea inițială sunt prezentate criteriile și terminologia de bază utilizate în dezvoltarea unui corpus, după care sunt introduse originile corpusurilor, fiind exemplificate primele corpusuri apărute. Corpusurile moderne sunt introduse împreună cu o prezentare generală a principalelor tipuri de corpusuri. Adnotarea și modalitățile de adnotare sunt tratate pe larg, acestea contribuind la dezvoltarea plajei de utilizări a corpusurilor. În plus, sunt discutate rolul corpusurilor în lingvistica computațională și posibilele utilizări ale acestora.

Capitolul 3 prezintă etapele premergătoare oricărui tip de procesare avansată a limbajului natural. Acestea având un rol foarte important în procesările ulterioare, influențează în mod direct performanța sistemelor de extragere a informațiilor din texte.

Capitolul 4 prezintă modalități de reprezentare a informațiilor în limbajul natural. Pentru exemplificare au fost alese două dintre cele mai utilizate resurse din acest domeniu, care au fost exploatate și în experimentele făcute în cadrul tezei, WordNet și SNOMED CT.

Capitolul 5 prezintă tipurile de învățare automată utilizate în procesarea limbajului natural.

Capitolul 6 prezintă recunoașterea entităților denumite, nivel teoretic, aceasta este una dintre principalele ramuri ale extragerii de informații din texte, cu ajutorul căreia se fac identificarea și clasificarea entităților denumite.

Capitolul 7 prezintă modalitățile de reprezentare vectorială a contextelor de utilizare a cuvintelor. Aceasta find una dintre cele mai de succes idei ale procesării moderne a limbajului natural, care a contribuit la dezvoltarea și îmbunătățirea a numeroase sisteme de extragere a informațiilor. În acest capitol sunt prezentate din punct de vedere teoretic cele două modele utilizate în generarea acestor tipuri de vectori, Skip-gram și CBOW.

Capitolul 8, care este cel mai cuprinzător capitol al tezei, prezintă implementările obiectivelor propuse, dar și progresele făcute în extragerea de informații în domeniul biomedical în limba română, acestea contribuind la deschiderea de noi orizonturi de cercetare în acest domeniul.

Concluzii și direcții viitoare

2.1 Contribuții

În conformitate cu obiectivele propuse, principalele contribuții ale tezei sunt următoarele:

1. Ca punct de plecare pentru cercetările viitoare au fost prezentate în detaliu: metodologia de creare a unui corpus, clasificarea corpusurilor existente, accentul fiind pus pe cele specializate, principalele metode utilizate în extragerea de informații din texte, schemele de adnotare folosite în adnotarea corpusurilor biomedical, dar și resursele existente utilizate în evaluarea sistemelor de PLN în domeniul biomedical.

2. A fost creată o resursă lingvistică unică pentru limba română, corpusul BioRo, cu scopul de a deveni un corpus de referință în limba română pentru limbajul biomedical, respectând cele mai bune standarde ale domeniului.

3. A fost creat corpusul MoNERo și a fost pus la dispoziția comunității științifice. Acesta este primul corpus „gold standard" biomedical în limba română adnotat la nivel morfologic și cu patru clase de entități denumite. Utilitatea acestui corpus a fost dovedită chiar în această teză, corpusul contribuind la adaptarea sistemelor de recunoaștere a entităților denumite la domeniul biomedical. În procesul de construire a acestui corpus a fost adoptată și o metodologie de adnotare a entităților denumite.

4. Pe baza corpusului BioRo au fost calculați vectorii semantici ai entităților denumite, țintă în cercetarea noastră, aceștia urmând a fi puși la dispoziția comunității de cerce-

tare. Importanța acestora reiese din rezultatele obținute în antrenarea sistemului de recunoaștere a entităților denumite utilizat, performanța acestuia fiind îmbunătățită.

5. Au fost testate două abordări de etichetare a entităților denumite.

2.2 Direcții viitoare

1. Îmbogăţirea corpusului BioRo cu texte din alte subdomenii medicale (genetică, pediatrie etc.) și adnotarea acestora cu entități denumite.
2. Crearea corpusurilor biomedicale în funcție de domenii (cardiologie, genetică etc.).
3. Adăugarea unui nou nivel de adnotare (sintactică) și a relațiilor semantice dintre concepte în corpusul MoNERo.
4. Dezvoltarea și îmbunătățirea performanței sistemelor de extragere a informațiilor din texte biomedicale.
5. Dezvoltarea unui set de test bilingv (română-engleză) pentru testarea similarității vectorilor de cuvinte, asemenea celui introdus pentru limba română de [187], dar adaptat pentru domeniul biomedical.
6. Introducerea în WordNet a termenilor medicali identificați ca entități denumite. În această direcție a fost făcut un studiu pentru a dezvolta metodologia de lucru [188].

2.3 Lucrări publicate

1. Andrei Coman, **Maria Mitrofan**, Dan Tufiș: Automatic identification and classification of legal terms in Romanian law texts, In press, 2019, ConsILR, Cluj, România.

2. Ioana Marinescu, Verginica Barbu Mititelu, **Maria Mitrofan**: Polishing MoNERo, the morphologically and medical named entities annotated corpus of Romanian, In press, 2019, ConsILR, Cluj, România.

3. Daniel Gîfu, Alex Moruz, Cecilia Bolea, Anca Bibiri, **Maria Mitrofan**: The methodology of building CoRoLa. On design, creation and use of of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo. Revue roumaine de linguistique, No./Issue 2, 2019 (LXIV).

4. Dan Tufiș, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, **Maria Mitrofan**, Mihaela Onofrei: Little strokes fell great oaks. Creating CoRoLa, the reference corpus of contemporary Romanian. On design, creation and use of of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo. Revue roumaine de linguistique, No./Issue 2, 2019 (LXIV).

5. Radu Ion, Vasile Păiș, **Maria Mitrofan**: RACAI's System at PharmaCoNER 2019. Proceedings of the PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track, EMNLP, 2019, Hong Kong, China.

6. **Maria Mitrofan**, Verginica Barbu Mititelu, Grigorina Mitrofan: MoNERo: a Biomedical Gold Standard Corpus for the Romanian Language. Proceedings of 18th ACL Workshop on Biomedical Natural Language Processing, ACL 2019, Florenţa, Italia.

7. Verginica Barbu Mititelu, Ivelina Stoyanova, Tsvetana Dimitrova, Svetlozara Leseva, **Maria Mitrofan**, Maria Todorova: Hear about Verbal Multiword Expressions in the Bulgarian and the Romanian Wordnets Straight from the Horse's Mouth. Proceedings of Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), ACL 2019, Florenţa, Italia.

8. Verginica Mititelu, **Maria Mitrofan**: Leaving No Stone Unturned When Identifying and Classifying Verbal Multiword Expressions in the Romanian Wordnet. Proceedings of the 10th Global WordNet Conference, GWC 2019, Wroclaw, Polonia

9. Elena Irimia, **Maria Mitrofan**, Verginica Mititelu: Evaluating the Wordnet and CoRoLa-based Word Embedding Vectors for Romanian as Resources in the Task of Microworlds Lexicon Expansion. Proceedings of the 10th Global WordNet Conference, GWC 2019, Wroclaw, Polonia

10. Dan Tufis, Verginica Barbu Mititelu, Elena Irimia, **Maria Mitrofan**, Radu Ion, George Cioroiu: Making Pepper Understand and Respond in Romanian. Proceedings of the 2019 22nd International Conference on Control Systems and Computer Science (CSCS), pp. 682-688. IEEE, 2019.

11. **Maria Mitrofan**, Verginica Barbu Mititelu, Grigorina Mitrofan: Towards the Construction of a Gold Standard Biomedical Corpus for the Romanian Language. Proceedings of MEDA „2nd Workshop on Curative Power of MEdical DAta", JCDL 2018, Fort Worth, Texas, SUA.

12. **Maria Mitrofan**, Dan Tufiş: BioRo: The Biomedical Corpus for the Romanian Language. Proceedings of Language Resources and Evaluation, LREC 2018, Miyazaki, Japonia.

13. **Maria Mitrofan**, Verginica Barbu Mititelu, Grigorina Mitrofan: A Pilot Study for Enriching the Romanian WordNet with Medical Terms. Proceedings of Computational Linguistics in Bulgaria, CLIB 2018, Sofia, Bulgaria.

14. **Maria Mitrofan**: Bootstrapping a Romanian Corpus for Medical Named Entity Recognition. Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria.

15. **Maria Mitrofan**, Radu Ion: Adapting the TTL Romanian POS Tagger to the Biomedical Domain. Proceedings of the Biomedical NLP Workshop associated with RANLP, 2017, Varna, Bulgaria.

16. **Maria Mitrofan** and Dan Tufiş Building and Evaluating the Romanian Medical Corpus. Proceedings of the 12 th International Conference „Linguistic Resources and tools for processing the Romanian language.", 2016, ConsILR, Iaşi, Romania.

Bibliografie

[1] Beth M Sundheim. Overview of the fourth message understanding evaluation and conference. In *Proceedings of the 4th conference on Message understanding*, pages 3–21, 1992.

[2] Berry De Bruijn and Joel Martin. Getting to the (c)ore of knowledge: mining biomedical literature. *International journal of medical informatics*, 67(1-3):7–18, 2002.

[3] Hagit Shatkay and Ronen Feldman. Mining the biomedical literature in the genomic era: an overview. *Journal of computational biology*, 10(6):821–855, 2003.

[4] Verginica Barbu Mititelu, Dan Tufiș, and Elena Irimia. The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

[5] P G W Glare. *Oxford Latin Dictionary*. Oxford, 1968.

[6] Philological Society (Great Britain). *Transactions of the Philological Society*. Society, 1870.

[7] Jan MG Aarts and Willem Meijs. *Corpus linguistics: Recent developments in the use of computer corpora in English language research*. Rodopi, 1984.

[8] John Sinclair. Corpus and text-basic principles. *Developing linguistic corpora: A guide to good practice*, pages 1–16, 2005.

[9] Anne O'Keeffe and Michael McCarthy. *The Routledge handbook of corpus linguistics*. Routledge, 2010.

[10] Roberto Busa. *Index Thomisticus Sancti Thomae Aquinatis Operum Omnium Indices Et Concordantiae in Quibus Verborum Omnium Et Singulorum Formae Et Lemmata Cum Suis Frequentiis Et Contextibus Variis Modis Referuntur*. 1974.

[11] Randolph Quirk. Towards a description of english usage. *Transactions of the philological society*, 59(1):40–61, 1960.

[12] W Nelson Francis and Henry Kucera. Manual of information to accompany a standard corpus of present-day edited american english, for use with digital computers. *Department of Linguistics, Brown University*, 1, 1964.

[13] Henry Kučera and Winthrop Nelson Francis. *Computational analysis of present-day American English*. Dartmouth Publishing Group, 1967.

[14] Stig Johansson, Geoffrey N Leech, and Helen Goodluck. The lancaster-oslo/bergen corpus of british english. *Department of English: Oslo UP*, 1978.

[15] Srikant V Shastri. The kolhapur corpus of indian english and work done on its basis so far. *ICAME Journal*, 12:15–26, 1988.

[16] John Sinclair. *Corpus, concordance, collocation.* Oxford University Press, 1991.

[17] BNC. Consortium et al. The british national corpus, version 3 (bnc xml edition). 2007. *Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www. natcorp. ox. ac. uk (last accessed 25th May 2012)*, 2012.

[18] Tim Berners-Lee. Long live the web. *Scientific American*, 303(6):80–85, 2010.

[19] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54, 2008.

[20] Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. The german reference corpus dereko: A primordial sample for linguistic research. In *Proceedings of the seventh international conference on Language Resources and Evaluation*, 2010.

[21] Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. The bulgarian national corpus: Theory and practice in corpus design. *Journal of Language Modelling*, (1):65–110, 2012.

[22] Marko Tadić. Building the croatian national corpus. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, 2002.

[23] Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, et al. Syn2015: representative corpus of contemporary written czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2522–2528, 2016.

[24] Csaba Oravecz, Tamás Váradi, and Bálint Sass. The hungarian gigaword corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2014.

[25] Yesim Aksan, Mustafa Aksan, Ahmet Koltuksuz, Taner Sezer, Ümit Mersinli, Umut Ufuk Demirhan, Hakan Yilmazer, Gülsüm Atasoy, Seda Öz, Ipek Yildiz, et al. Construction of the turkish national corpus (tnc). pages 3223–3227, 2012.

[26] Douglas Biber, Susan Conrad, and Randi Reppen. *Corpus linguistics: Investigating language structure and use.* Cambridge University Press, 1998.

[27] Douglas Biber. Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257, 1993.

[28] Jan Pomikálek, Milos Jakubícek, and Pavel Rychlỳ. Building a 70 billion word corpus of english from clueweb. In *LREC*, pages 502–506, 2012.

[29] Tony McEnery, Andrew Wilson, and Andrew Wilson. *Corpus linguistics: An introduction*. Edinburgh University Press Edinburgh, 2001.

[30] Brian Clancy. Building a corpus to represent a variety of a language. In *The Routledge handbook of corpus linguistics*, pages 108–120. Routledge, 2010.

[31] Douglas Biber. *Variation across speech and writing*. Cambridge University Press, 1991.

[32] Graeme Kennedy. *An introduction to corpus linguistics*. Routledge, 2014.

[33] Almut Koester. Building small specialised corpora. In *The Routledge handbook of corpus linguistics*, pages 94–107. Routledge, 2010.

[34] Michael Stubbs. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of language*, 2(1):23–55, 1995.

[35] Tony McEnery, Richard Xiao, and Yukio Tono. *Corpus-based language studies: An advanced resource book*. Taylor & Francis, 2006.

[36] Paul Baker. *Glossary of Corpus Linguistics*. Edinburgh University Press, 2006.

[37] J Sinclair and J Ball. Eagles text typology. *Internal Working Document*, 1995.

[38] Jaroslav Blecha. *Building Specialized Corpora*. PhD thesis, Masarykova univerzita, Filozofická fakulta, 2013.

[39] Dan Cristea. Resurse lingvistice şi tehnologiile limbajului natural. cazul limbii române. *Prelegeri Academice*, III(3), 2012. ISSN 1583-4514.

[40] Jennifer Pearson. *Terms in Context*, volume 14. John Benjamins Publishing, 1998.

[41] J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. Genia corpus a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(1):180–182, 2003.

[42] JB Carroll, P Davies, and B Richman. The american heritage intermediate corpus. In *Proceedings of the International Conference on Computational Linguistics*. New York: American Heritage Publishing Co, 1971.

[43] Konrad Hofbauer, Stefan Petrik, and Horst Hering. The atcosim corpus of non-prompted clean air traffic control speech. In *Proceedings of the 6th edition of the Language Resources and Evaluation Conference.*, 2008.

[44] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

[45] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufiş, and Dániel Varga. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 2006.

[46] Sidney Greenbaum. The development of the international corpus of english. In *English corpus linguistics*, pages 95–104. Routledge, 2014.

[47] Inguna Skadiņa, Andrejs Vasiļjevs, Raivis Skadiņš, Robert Gaizauskas, Dan Tufiş, and Tatiana Gornostay. Analysis and evaluation of comparable corpora for under-resourced areas of machine translation. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, pages 17–19, 2010.

[48] Matti Rissanen, M Kytö, L Kahlas-Tarkka, M Kilpiö, S Nevanlinna, I Taavitsainen, T Nevalainen, and H Raumolin-Brunberg. The helsinki corpus of english texts. *Department of English University of Helsinki*, 1993.

[49] Marianne Hundt, Andrea Sand, and Paul Skandera. *Manual of Information to Accompany the Freiburg-Brown Corpus of American English ('Frown')*. Albert-Ludwigs-Universität Freiburg, 1999.

[50] Mark Davies. Diachronic corpus of present-day spoken english (dcpse). 2009.

[51] Jan Svartvik. *The London-Lund corpus of spoken English: Description and research*. Number 82. Lund University Press, 1990.

[52] Catalina Maranduc. A diachronic corpus for romanian (rodia). *Proceedings of the LT4DHCSEE in conjunction with RANLP*, pages 1–9, 2017.

[53] Lynne Bowker and Jennifer Pearson. *Working with specialized language: a practical guide to using corpora*. Routledge, 2002.

[54] Antoinette Renouf. Corpus development 25 years on: from super-corpus to cybercorpus. *Language and computers studies in practical linguistics*, 62(1):27, 2007.

[55] T-WILSON McENERY and A Wilson. A.(1996) corpus linguistics, 2001.

[56] Ruslan Mitkov. *The Oxford handbook of computational linguistics*. Oxford University Press, 2005.

[57] Geoffrey Leech. Introducing corpus annotation. *Corpus Annotation–Linguistic Information from Computer Text Corpora*, pages 1–18, 1997.

[58] John M Sinclair. The automatic analysis of corpora. *Directions in Corpus Linguistics*, 65:379–397, 1992.

[59] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19 (2):313–330, 1993.

[60] Atro Voutilainen and Timo Järvinen. Specifying a shallow grammatical representation for parsing purposes. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pages 210–214. Morgan Kaufmann Publishers Inc., 1995.

[61] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics*, volume 1, pages 86–90. Association for Computational Linguistics, 1998.

[62] Simon Philip Botley and Tony McEnery. *Corpus-based and computational approaches to discourse anaphora*, volume 3. John Benjamins Publishing, 2000.

[63] Anna Brita Stenström. *Questions and responses in English conversation*. Krieger Pub Co, 1984.

[64] Sidney Greenbaum and Jan Svartvik. *The london-lund corpus of spoken english*, volume 7. Lund University Press, 1990.

[65] Béatrice Daille. Combined approach for terminology extraction: Lexical statistics and linguistic filtering. Citeseer, 1995.

[66] Ezra Black, Roger Garside, and Georey Leech. Statistically-driven computer grammars of english: The ibm/lancaster approach. *Lancaster Approach, Amsterdam*, 1993.

[67] Tomaž Erjavec and Nancy Ide. The multext-east corpus. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 971–74. Citeseer, 1998.

[68] Holger Schwenk and Jean-Luc Gauvain. Training neural network language models on very large corpora. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 201–208, 2005.

[69] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[70] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*, 2013.

[71] Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Automatic extraction of rules for sentence boundary disambiguation. In *Proceedings of the Workshop on Machine Learning in Human Language Technology*, pages 88–92, 1999.

[72] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[73] Radu Ion. *Word sense disambiguation methods applied to English and Romanian*. PhD thesis, Romanian Academy, Bucharest, 2007.

[74] Songjian Chen, Yabo Xu, and Huiyou Chang. A simple and effective unsupervised word segmentation approach. In *Proceedings of Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[75] Mark Johnson. Unsupervised word segmentation for sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, 2008.

[76] Thorsten Brants. Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231. Association for Computational Linguistics, 2000.

[77] Christer Samuelsson. Morphological tagging based entirely on bayesian inference. In *Proceedings of the 9th Nordic Conference of Computational Linguistics (NODALIDA)*, pages 225–238, 1994.

[78] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, 2001.

[79] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1996.

[80] Nuno C Marques, Gabriel Pereira Lopes, et al. A neural network approach to part-of-speech tagging. In *Proceedings of the 2nd meeting for computational processing of spoken and written Portuguese*, pages 21–22, 1996.

[81] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[82] Michael Collins, Lance Ramshaw, Jan Hajič, and Christoph Tillmann. A statistical parser for czech. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 505–512. Association for Computational Linguistics, 1999.

[83] D Tufiş, AM Barbu, V Pătraşcu, G Rotariu, and C Popescu. Corpora and corpus-based morpho-lexical processing. In *Proceedings of the Recent Advances in Romanian Language Technology*, pages 35–56. Editura Academiei, 1997.

[84] Erjavec Tomaž. Multext-east and tei: an investigation of a schema for language engineering and corpus linguistics. *Multilinguality and interoperability in language processing with emphasis on Romanian*, pages 19–47, 2010.

[85] J.P. Ferraro, H. Daumé III, S. L. DuVall, W. W. Chapman, H. Harkema, and P. J. Haug. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *Journal of the American Medical Informatics Association*, 20(5):931–939, 2013.

[86] Dan Tufiş. Tiered tagging and combined language models classifiers. In *Proceedings of the International Workshop on Text, Speech and Dialogue*, pages 28–33. Springer, 1999.

[87] Alexandru Ceauşu. Maximum entropy tiered tagging. In *Proceedings of the 11th ESSLLI student session*, pages 173–179. Citeseer, 2006.

[88] CJ Van Rijsbergen. *Information retrieval.* University of Glasgow, 1979.

[89] Philip Resnik and David Yarowsky. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural language engineering*, 5(2):113–133, 1999.

[90] Tom Gruber. Ontology. *Encyclopedia of Database Systems*, 2008.

[91] Natalya F Noy, Deborah L McGuinness, et al. Ontology development 101: A guide to creating your first ontology. *Stanford knowledge systems laboratory technical report*, 2001.

[92] George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.

[93] Piek Vossen. A multilingual database with lexical semantic networks. *Computers and the Humanities*, 10:978–994, 1998.

[94] Sofia Stamou, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufiş, Svetla Koeva, George Totkov, Dominique Dutoit, and Maria Grigoriadou. Balkanet: A multilingual semantic network for the balkan languages. In *Proceedings of the 1st International Wordnet Conference*, pages 21–25, 2002.

[95] Dan Tufiş. Ro-wordnet: ontologie lexicală pentru limba română. *Academica*, 18 (208-209):30–34, 2008.

[96] Barry Smith and Christiane Fellbaum. Medical wordnet: a new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 371–382, 2004.

[97] Stephen Marsland. *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC, 2011.

[98] Tom M Mitchell et al. Machine learning. 1997. *Annual review of computer science*, 45(37):870–877, 1997.

[99] Claire Cardie and Raymond J Mooney. Guest editors' introduction: Machine learning and natural language. *Machine Learning*, 34(1):5–9, 1999.

[100] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[101] Alex Graves. Supervised sequence labelling. In *Supervised sequence labelling with recurrent neural networks*, pages 5–13. Springer, 2012.

[102] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.

[103] G David Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[104] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[105] Alex Graves, Abdel Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp)*, pages 6645–6649. IEEE, 2013.

[106] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[107] Vijay Patil and Sanjay Shimpi. Handwritten english character recognition using neural network. *Elixir Comput Sci Eng*, 41:5587–5591, 2011.

[108] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[109] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[110] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[111] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[112] Daniel Jurafsky and James H Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. *Prentice-Hall*, 2000.

[113] Lisa F Rau. Extracting company names from text. In *Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications*, volume 1, pages 29–32, 1991.

[114] Changki Lee, Yi-Gyu Hwang, Hyo-Jung Oh, Soojong Lim, Jeong Heo, Chung-Hee Lee, Hyeon-Jin Kim, Ji-Hyun Wang, and Myung-Gil Jang. Fine-grained named entity recognition using conditional random fields for question answering. In *Proceedings of the Asia Information Retrieval Symposium*, pages 581–587, 2006.

[115] Javier Artiles, Julio Gonzalo, and Satoshi Sekine. Weps 2 evaluation campaign: overview of the web people search clustering task. In *Proceedings of the 2nd web people search evaluation workshop (WePS)*, volume 9, 2009. URL http://www2009.eprints.org/257/.

[116] Mijail Kabadjov, Josef Steinberger, and Ralf Steinberger. Multilingual statistical news summarization. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 229–252. Springer, 2013.

[117] DANIELA Gîfu and GABRIELA Vasilache. A language independent named entity recognition system. *Alexandru Ioan Cuza" University Publishing House, Iaşi*, pages 181–188, 2014.

[118] Thierry Poibeau and Leila Kosseim. Proper name extraction from non-journalistic texts. *Language and computers*, 37:144–157, 2001.

[119] Massimiliano Ciaramita and Yasemin Altun. Named-entity recognition in novel domains with external lexical knowledge. In *Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*, 2005.

[120] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In *Proceedings of the The Third International Conference on Language Resources and Evaluation*, 2002.

[121] Jenny Rose Finkel and Christopher D Manning. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610. Association for Computational Linguistics, 2009.

[122] Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a high-performance learning name-finder. In *Proceedings of the fifth conference on Applied natural language processing*, pages 194–201. Association for Computational Linguistics, 1997.

[123] Michael Collins and Yoram Singer. Unsupervised models for named entity classification. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[124] Erik F Sang and Jorn Veenstra. Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 173–179, 1999.

[125] Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 359–367, 2011.

[126] Jana Straková, Milan Straka, and Jan Hajič. A new state-of-the-art czech named entity recognizer. In *Proceedings of the International Conference on Text, Speech and Dialogue*, pages 68–75. Springer, 2013.

[127] Younggyun Hahm, Jungyeul Park, Kyungtae Lim, Youngsik Kim, Dosam Hwang, and Key-Sun Choi. Named entity corpus construction using wikipedia and dbpedia ontology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 2565–2569, 2014.

[128] Joohui An, Seungwoo Lee, and Gary Geunbae Lee. Automatic acquisition of named entity tagged corpus from world wide web. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*, pages 165–168. Association for Computational Linguistics, 2003.

[129] Beth M Sundheim. Overview of results of the muc-6 evaluation. In *Proceedings of the 6th conference on Message understanding*, pages 13–31. Association for Computational Linguistics, 1995.

[130] Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, volume 4, pages 142–147, 2003.

[131] Alexander Tkachenko, Timo Petmanson, and Sven Laur. Named entity recognition in estonian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 78–83, 2013.

[132] Hakan Demir and Arzucan Ozgur. Improving named entity recognition for morphologically rich languages using word embeddings. In *Proceedings of the 13th International Conference on Machine Learning and Applications*, pages 117–122, 2014.

[133] Georgi Georgiev, Preslav Nakov, Kuzman Ganchev, Petya Osenova, and Kiril Simov. Feature-rich named entity recognition for bulgarian using conditional random fields. In *Proceedings of the International Conference RANLP-2009*, pages 113–117, 2009.

[134] Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1002–1012. Association for Computational Linguistics, 2010.

[135] GuoDong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480, 2002.

[136] Vincent Labatut. Improved named entity recognition through svm-based combination. *HAL Archives-Ouvertes*, 2013.

[137] Xiaohua Liu, Furu Wei, Shaodian Zhang, and Ming Zhou. Named entity recognition for tweets. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(1):3, 2013.

[138] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[139] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):37–48, 2017.

[140] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.*, pages 2741–2749, 2016.

[141] Thai-Hoang Pham and Phuong Le-Hong. End-to-end recurrent neural network models for vietnamese named entity recognition: Word-level vs. character-level. In *Proceedings of the International Conference of the Pacific Association for Computational Linguistics*, pages 219–232, 2017.

[142] Onur Kuru, Ozan Arkan Can, and Deniz Yuret. Charner: Character-level named entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 911–921, 2016.

[143] Rodrigo Agerri and German Rigau. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82, 2016.

[144] David Nadeau, Peter D Turney, and Stan Matwin. Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In *Proceedings of the Conference of the Canadian Society for Computational Studies of Intelligence*, pages 266–277. Springer, 2006.

[145] Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

[146] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[147] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2016.

[148] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[149] JR Firth. *Papers in Linguistics*. Oxford University Press, 1957.

[150] Andrew Trask, Phil Michalak, and John Liu. sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. *CoRR*, 2015.

[151] Jun-Tae Kim and Dan I. Moldovan. Acquisition of linguistic patterns for knowledge-based information extraction. *IEEE transactions on knowledge and data engineering*, 7(5):713–724, 1995.

[152] Stephen Soderland. Learning information extraction rules for semi-structured and free text. *Machine learning*, 34(1-3):233–272, 1999.

[153] Mary Elaine Califf and Raymond J Mooney. Bottom-up relational learning of pattern matching rules for information extraction. *Journal of Machine Learning Research*, 4 (Jun):177–210, 2003.

[154] Maria Mitrofan and Dan Tufis. Bioro: The biomedical corpus for the romanian language. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, 2018.

[155] John Sinclair. Trust the text. In *Advances in written text analysis*, pages 26–39. Routledge, 2002.

[156] David Lee and John Swales. A corpus-based eap course for nns doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for specific purposes*, 25(1):56–75, 2006.

[157] Ching-Fen Chang and Chih-Hua Kuo. A corpus-based approach to online materials development for writing research articles. *English for Specific Purposes*, 30(3):222–234, 2011.

[158] Paul Thompson and Chris Tribble. Looking at citations: Using corpora in english for academic purposes. *Language learning and technology*, 5(3):91–105, 2001.

[159] Thomas A Upton and Ulla Connor. Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes*, 20(4): 313–329, 2001.

[160] Lynne Flowerdew. The argument for using english specialized corpora to understand academic and professional language. *Discourse in the professions: Perspectives from corpus linguistics*, pages 11–33, 2004.

[161] Timothy D Imler, Justin Morea, and Thomas F Imperiale. Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals. *Clinical Gastroenterology and Hepatology*, 12(7):1130–1136, 2014.

[162] Joshua C Denny, Anderson Spickard III, Peter J Speltz, Renee Porier, Donna E Rosenstiel, and James S Powers. Using natural language processing to provide personalized learning opportunities from trainee clinical notes. *Journal of biomedical informatics*, 56:292–299, 2015.

[163] Wendy W Chapman, Adi V Gundlapalli, Brett R South, and John N Dowling. Natural language processing for biosurveillance. In *Infectious Disease Informatics and Biosurveillance*, pages 279–310. Springer, 2011.

[164] Katherine P Liao, Tianxi Cai, Guergana K Savova, Shawn N Murphy, Elizabeth W Karlson, Ashwin N Ananthakrishnan, Vivian S Gainer, Stanley Y Shaw, Zongqi Xia, and Peter Szolovits. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*, 350, 2015.

[165] Pierre Zweigenbaum, Robert Baud, Anita Burgun, Fiammetta Namer, Éric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-François Forget, Magaly Douyere, et al. Umlf: a unified medical lexicon for french. *International Journal of Medical Informatics*, 74(2-4):119–124, 2005.

[166] Svetla Boytcheva, Ivelina Nikolova, Elena Paskaleva, Galia Angelova, Dimitar Tcharaktchiev, and Nadya Dimitrova. Extraction and exploration of correlations in patient status data. In *Proceedings of the Workshop on Biomedical Information Extraction*, pages 1–7. Association for Computational Linguistics, 2009.

[167] Sumithra Velupillai. *Shades of certainty: annotation and classification of swedish medical records*. PhD thesis, Department of Computer and Systems Sciences, Stockholm University, 2012.

[168] Alex Moruz and Andrei Scutelnicu. An automatic system for improving boilerplate removal for romanian texts. In *Proceedings of the 10th International Conference "Linguistic resources and Tools for Processing the Romanian Language*, pages 163–170, 2014.

[169] Piotr Bański, Nils Diewald, Michael Hanl, Marc Kupietz, and Andreas Witt. Access control by query rewriting: the case of korap. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, page 3817–3822, 2014.

[170] Nils Diewald, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Banski, and Andreas Witt. Korap architecture-diving in the deep sea of corpus data. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, page 3586–3591, 2016.

[171] Y. Tsuruoka, Y.Tateishi, J. D. Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. I. Tsujii. Developing a robust part-of-speech tagger for biomedical text. *Advances in informatics*, pages 382–392, 2005.

[172] Vimla L Patel, Edward H Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. The coming of age of artificial intelligence in medicine. *Artificial intelligence in medicine*, 46(1):5–17, 2009.

[173] Adam S Rothschild, Harold P Lehmann, and George Hripcsak. Inter-rater agreement in physician-coded problem lists. In *Proceedings of the AMIA Annual Symposium*, volume 2005, pages 644–648. American Medical Informatics Association, 2005.

[174] Yasunori Yamamoto, Atsuko Yamaguchi, Hidemasa Bono, and Toshihisa Takagi. Allie: a database and a search service of abbreviations and long forms. *Database*, 2011, 2011.

[175] Hongfang Liu, Alan R Aronson, and Carol Friedman. A study of abbreviations in medline abstracts. In *Proceedings of the AMIA Symposium*, pages 464–468, 2002.

[176] Baohua Gu. Recognizing nested named entities in genia corpus. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, pages 112–113, 2006.

[177] Matthew Lease and Eugene Charniak. Parsing biomedical literature. In *International Conference on Natural Language Processing*, pages 58–69. Springer, 2005.

[178] Parikshit Sondhi. A survey on named entity extraction in the biomedical domain. Department of Computer Science University of Illinois, 2008.

[179] Tomaž Erjavec. Multext-east: morphosyntactic resources for central and eastern european languages, language resources and evaluation. *Language resources and evaluation*, 46:131–142, 2012.

[180] Pătraşcu V. Rotariu G. Popescu C. Dan Tufiş, Barbu A.M. Corpora and corpus-based morpho-lexical processing. *Recent Advances in Romanian Language Technology*, pages 35–56, 1997.

[181] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, pages 249–254, 1996.

[182] Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton. Semantic similarity and relatedness between clinical terms: an experimental study. In *Proceedings of the AMIA annual symposium*, pages 572–576, 2010.

[183] Maria Mitrofan. Bootstrapping a romanian corpus for medical named entity recognition. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 501–509, 2017.

[184] Tiberiu Boroș and Ruxandra Burtica. Gbd-ner at parseme shared task 2018: Multi-word expression detection using bidirectional long-short-term memory networks and graph-based decoding. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, pages 254–260, 2018.

[185] Eliyahu Kiperwasser and Yoav Goldberg. Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 2016.

[186] Vasile Paiş and Dan Tufiş. Computing distributed representations of words using the corola corpus. *Proceedings of the Romanian Academy Series A Mathematics Physics Technical Sciences Information Science*, 19(2):403–409, 2018.

[187] Samer Hassan and Rada Mihalcea. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 3, pages 1192–1201, 2009.

[188] Maria Mitrofan, Verginica Barbu Mititelu, and Grigorina Mitrofan. A pilot study for enriching the romanian wordnet with medical terms. In *Proceedings of the Third International Conference Computational Linguistics in Bulgaria*, 2018.