Annoteringar gjorda 2008-2023

För flera av nedanstående annoteringsuppgifter har vi gjort olika testrundor och piloter som inte är medtagna i uppräkningen.

Stockholm EPR PHI Corpus

Annoterare: Gunnar Nilsson, Hercules Dalianis, Sumithra Velupillai, hösten 2008

100 patientjournaler, 5 kliniker, Neurologi, Ortopedi, Käkkirurgi, Infektion samt Dietistklinik.
20 patienter från varje klinik fördelade på 50 procent kvinnor och 50 procent män

Konsensus:

Age: 56 instanser
First_Name: 923 instanser
Last_Name: 929 instanser
Phone_Number: 137 instanser
Location: 148 instanser
Health_Care_Unit: 1025 instanser
Date_Part: 711 instanser
Full_Date: 551 instanser
Totalt: 4480 instanser, 380 000 tokens totalt (174 000 tokens om man tar bort tabellinformation)

Ligger på KEA-datorn: shared/annotations/Stockholm EPR PHI Corpus (8 PHI Classes)

Referenser

Velupillai, S., H. Dalianis, M. Hassel and G. H. Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. International Journal of Medical Informatics (2009), doi:10.1016/j.ijmedinf.2009.04.005

Dalianis, H. and S. Velupillai. 2010. De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields, Journal of Biomedical Semantics 2010, 1:6 (12 April 2010)

Carlsson, E. and H. Dalianis. 2010. Influence of Module Order on Rule-Based De-identification of Personal Names in Electronic Patient Records Written in Swedish, in the Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, May 19-21, 2010, pp 3071-3075

Stockholm EPR PHI Pseudo Corpus ver1

Stockholm EPR PHI Pseudo Corpus finns som psudonymiserad korpus, arbetet beskrivs I Alfalahi et al (2012)

Ligger på KEA-datorn: shared/annotations/Pseudo-Stockholm EPR PHI Pseudo Corpus (8 PHI Classes) ver1

Referenser

Alfalahi, A., S. Brissman and H. Dalianis. 2012. Pseudonymisation of person names and other PHIs in an annotated clinical Swedish corpus. In the Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, May 26, Istanbul, pp 49-54.

Stockholm EPR PHI Pseudo Corpus ver3

Stockholm EPR PHI Pseudo Corpus finns som psudonymiserad korpus, arbetet beskrivs I Dalianis 2010.

Ligger på KEA-datorn: shared/annotations/Pseudo-Stockholm EPR PHI Pseudo Corpus (8 PHI Classes) ver3

Referenser

Dalianis, H. 2019. Pseudonymisation of Swedish Electronic Patient Records Using a Rule-based Approach. In the Proceedings of the Workshop on NLP and Pseudonymisation, in conjunction with the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), Turku, Finland, September 30, 2019

Berg, H., T. Chomutare and H. Dalianis. 2019. Building a De-identification System for Real Swedish Clinical Text Using Pseudonymised Clinical Text. In the Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis, Louhi 2019, in conjuction with Conference on Empirical Methods in Natural Language Processing, (EMNLP) November 2019, Hongkong, ACL, pp 118-125.

Stockholm EPR PHI Domain Corpus ver 1

Annoterare: Mia Kvist, hösten 2016

3 kliniker, Geriatrik, Onkology och Ortopedi inkluderat kirurgi.

Konsensus:

Age: 36 instanser

First_Name: 251 instanser

Last_Name: 325 instanser

Phone_Number: 18 instanser

Location: 67 instanser

Health_Care_Unit: 330 instanser

Date_Part: 337 instanser

Full_Date: 215 instanser

Totalt: 1 579 instanser, 63 417 tokens totalt (174 000 tokens om man tar bort tabellinformation)

Ligger på KEA shared/annotations/Stockholm EPR PHI Domain Corpus ver 1

Referenser

Henriksson, A., M. Kvist, and H. Dalianis. 2017a. Prevalence Estimation of Protected Health

Information in Swedish Clinical Text, *Informatics for Health, Manchester, UK, 24-26 April, 2017*

Henriksson A., M. Kvist, and H. Dalianis. 2017b, Detecting Protected Health Information in Heterogeneous Clinical Notes, *MedInfo, Hangzhou, China, Aug 20-25, 2017*

Stockholm EPR PHI Domain Corpus ver 2

(Uppdaterad version av Stockholm EPR PHI Domain Corpus ver 2).
Annoterare: Hanna Berg, sommaren 2019

I den uppdaterade versionen:

Age: 21 instanser
First_Name: 172 instanser
Last_Name: 243 instanser
Phone_Number: 19 instanser
Location: 40 instanser
Health_Care_Unit: 179 instanser
Date_Part: 337 instanser
Full_Date: 215 instanser
Totalt: 1 079 instanser, 50 519 tokens totalt

Ändringar Hanna Berg:
- Lagt till förnamn som missats
- Lagt till hela datum som missats
- Datumdelar som bara består av år har blivit till O (skulle följa Stockholm EPR PHI)
- Tagit bort felaktiga locations (ex. augusti, flyktingförläggning)
- Bytt någon del locations till hcu (Furuhöjden, Huddinge när det avser sjukhuset)
- Generalla HCU -> O (ex VC, akuten)
- Lagt till ett missat telefonnummer

Det där är ju däremot inte hela domänkorpusen, då endast 50 519 tokens tokens var annoterade av 116 111 tokens, Hanna annoterade nedan också ”Other half”

Age: 34 instanser
First_Name: 208 instanser
Last_Name: 282 instanser
Location 57: instanser
Health_Care_Unit: 207 instanser
Phone_Number: 22 instanser
Full_Date: 225 instanser
Date_Part: 309 instanser
Totalt: 1 334 instanser, 65 592 tokens totalt
och för **båda** delarna 2 423 instanser totalt

Ligger på KEA shared/annotations/Stockholm EPR PHI Domain Corpus ver 2

Referenser

Berg, H. and H. Dalianis. 2019. Augmenting a De-identification System for Swedish Clinical Text Using Open Resources (and Deep learning). In the Proceedings of the Workshop on NLP and Pseudonymisation, in conjunction with the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), Turku, Finland, September 30, 2019.

NorDeid corpus

Baserad på Stockholm EPR PHI Pseudo Corpus
Balanserad korpora
Permutera slumpmässigt nya PHIs från samma resurs in den nya korpusen
För entiteter med frekvens mindre än 100 använd ScandiBERT att generera nya PHIs till corpusen (Fill in the gap)

	Pseudonymized Stockholm EPR PHI Corpus	Balanced corpora (Dataset II)
First Name	928	5,999
Last Name	923	5,170
Phone Number	135	4,514
Age	56	2,102
Full Date	500	4,373
Date Part	710	4,381
Health Care Unit	1,021	7,653
Location	95	4,355
Organisation	53	3,727
Total	4,421	42,274

Ligger på KEA shared/annotations/NorDeid-Corpus/

Referenser

Lamproudis, A., Mora, S., Olsen Svenning T., Torsvik T., Chomutare T., Budrionis A, Dinh Ngo P. and H. Dalianis. 2023. De-identifying Norwegian Clinical Text using Resources from Swedish and Danish, to appear in the Proceedings of AMIA 2023, Annual Symposium, November 11-15. New Orleans, LA, USA.

Stockholm EPR Sentence Uncertainty Corpus (sommarmängden)

Annoterare: Aron Henriksson, Helen Allvin, Freja Dalianis, Hercules Dalianis, Sumithra Velupillai, sommar 2009

6 740 slumpmässigt uttagna meningar ur hela Stockholm EPR Corpus, endast ur bedömningsfält.

Annoteringsklasser:

- Meningsnivå: certain expression, uncertain expression, undefined expression

”Ord”nivå: speculative words, negation, undefined speculative words
Certain_expression: 4 938 instanser
Uncertain_expression: 582 instanser

Undefined expression: 146 instanser

Speculative words: 1 077 instanser

Negation: 910 instanser

Undefined_speculative_words: 0 instanser

Summa: 7 653 instanser

Ligger på KEA-datorn: shared/annotations/Stockholm EPR Sentence Uncertainty Corpus (Summer 2009)

Referenser

Dalianis, H. and S. Velupillai. 2010. How Certain are Clinical Assessments? Annotating Swedish Clinical Text for (Un)certainties, Speculations and Negations. In Proceedings of the of the Seventh International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, May 19-21 2010. URL

Velupillai, S. 2010. Towards a better understanding of uncertainties and speculations in Swedish clinical text - analysis of an initial annotation trial. In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 14–22, Uppsala, Sweden, July 10 2010. University of Antwerp. ISBN 9789057282669.

Dalianis, H. and M. Skeppstedt. 2010. Creating and Evaluating a Consensus for Negated and Speculative Words in a Swedish Clinical Corpus. In the Proceedings of the Negation and Speculation in Natural Language Processing, NeSp-NLP 2010 Workshop, July 10, 2010, University of Antwerp, pp 5-13, Uppsala, Sweden, Publisher: ACL 2010

Stockholm EPR Diagnosis Factuality Corpus

Annoterare: Maria Kvist, Gunnar Nilsson, Sumithra Velupillai, Hercules Dalianis, hösten 2010

S_medakut, endast bedömningsfält. Batcher med mappar à 50 fält, diagnosuttryck förmarkerade automatiskt efter ovanstående lista + Granskatagger.

Annoteringsklasser:

Certainly_positive, Probably_positive, Possibly_positive, Possibly_negative, Probably_negative, Certainly_negative, Other, Not_diagnosis.

Temporality_past: användes enbart om default (nutid) ej var tillämpbart.

Antal instanser per klass (OBS Mias annoteringar batch 3 + 4 + 5)

Certainly Positive: 3 088
Probably Positive: 1 039
Possibly Positive: 663
Possibly Negative: 139
Probably Negative: 546
Certainly Negative: 711

Not Diagnosis: 117
Other: 180
Summa: 6 303 annoterade tokens av totalt 283 007 tokens

Ligger på KEA-datorn: shared/annotations/ Stockholm EPR Diagnosis Factuality Corpus

Guidelines:
http://www.dsv.su.se/hexanord/guidelines/guidelines_stockholm_epr_diagnosis_factuality_corpus.pdf

Batcher:

Batch 1 + 2: testbatcher, ej använda för analys.

Batch 3: annoterad av Mia + Gunnar: 26 mappar, totalt 1297 bedömningsfält. Antal instanser: 2 182 (2 070 analyserade för Inter-annotator agreement pga skillnader i användning av temporality_past).

Batch 3: även återskapad i ny slumpmässig ordning för Intra-annotator agreement (Mia)

Batch 4 + 5: annoterades av Mia, totalt 3 841 bedömningsfält, 6 483 instanser.

Batch 6 – 9: ej annoterade, men finns tillgängliga på KEA.

Övrigt:

- Täckningsanalys (batch 3):
  Mia taggar upp de diagnoser som ej var markerade i de 26 mappar som hade taggats automatiskt, samt även i de 24 mappar där ingen tagg hade fastnat vid den automatiska taggningen. Allt detta gjort 2 ggr, 2:a gången random, för att kolla IAA intra. Detta material har dock ännu ej använts i någon studie.
- pyConTextSwe: filer med triggrar, databas, pythonkod mm ligger på KEA1 under /data0/pyConTextSwe (OBS: själva pyConTextNLP finns kanske inte på KEA?)

Referenser:

Velupillai S., H. Dalianis, and M. Kvist. 2011. Factuality levels of diagnoses in Swedish clinical text. In Proceedings of MIE 2011, pages 559–563. IOS Press, August.

Velupillai S. 2011. Automatic Classification of Factuality Levels – A Case Study on Swedish Diagnoses and the Impact of Local Context. In Proc. The Fourth International Symposium on Languages in Biology and Medicine (LBM 2011), Singapore, December 2011.

Velupillai, S., and Kvist, M. 2012. Fine-Grained Certainty Level Annotations Used for Coarser-Grained E-Health Scenarios. In *Computational Linguistics and Intelligent Text Processing* (pp. 450-461). Springer Berlin Heidelberg.

Stockholm EPR Clinical Entity Corpus

Annoterare: Mia Kvist, Maria Skeppstedt, Gunnar Nilsson.

Annoterat Diagnos, Finding, Body part och Drug

Guidelines här: http://people.dsv.su.se/~mariask/resources/guidelines_entity_annotation.pdf

(Testannoteringar i batch 9:mapp 2-19, Med akuten)

- Medicinakuten
  Nedanstående annoterat av Mia, ett subset annoterat av Maria (batch 9 : note 20:1-20: mapp 42, mapp 21, mapp 22, 24, notes 30:1-25) samt av Gunnar (batch 9: mapp 23, 24, 25, 26, 27)
  batch 9: mappar 20-47 = 1313 bedömningsfält
  (miss när sparat gör att vissa fält ej sparats annoterade)
  Pga trassel med minnet blev 72 bedömningsfält annoterade dubbelt och kan användas för intra IAA: batch 9, fälten 32: 1-50 samt 33: 1-22.
  - disorders (1988 instances) (Diagnosis proper)
  - findings (3681 instances)
  - body structures (735 instances)
  - drugs (1 542 Instances, 482 types)

  Totalt 7 946 annoterade instanser 70 852 tokens

- Mapp 20-43 finns det även en annan version av som är jämförd mot maskininlärt data.
- ORT akuten (Ortopediakuten)
  Annoterat 20 högar á 50 = 1000 bedömningsfält
  - disorders (1258 instances, 541 types) (Diagnosis proper)
  - disorders implicit (10 instances, 7 types) (Diagnosis unproper)
  - findings (1439 instances, 785 types)
  - body structures (1324 instances, 423 types)
  - drugs (880 Instances, 212 types)
- HIA (Hjärtintensivenavdelningen)
  Annoterat x högar á 50 = y bedömningsfält
  - disorders (1088 instances, 533 types) (Diagnosis proper)
  - disorders implicit (20 instances, 18 types) (Diagnosis unproper)
  - findings (1798 instances, 1295 types)
  - body structures (461 instances, 252 types)
  - drugs (1 048 Instances, 497 types)

Ligger på KEA-datorn: shared/annotations/Stockholm EPR Clinical Entity Corpus

Referenser:

Skeppstedt, M., M. Kvist, H. Dalianis and Nilsson, G.H. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. Journal of Biomedical Informatics, DOI: 10.1016/j.jbi.2014.01.012

Skeppstedt, M., M. Kvist and H. Dalianis. 2012. Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. In the Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, May 23-25, Istanbul, pp 1250-125

Sidrat ul Muntaha, Maria Skeppstedt, Maria Kvist and Hercules Dalianis. 2012
Automatic Rule Based Detection of Pharmaceutical Drugs in Swedish Clinical Text.
In the proceedings of the Fourth Swedish Language Technology Conference, (SLTC-2012), Lund, Sweden, October 25-26, 2012

Annoterare: Gunnar Nilsson och Mia Kvist, Hösten 2010.

- Annotering för Diagnos, Finding, Diagnos & Body part, mm.
  10 mappar á 50 bedömningsfält (medicinakuten), men jag är osäker på om vi annoterade alla. Av detta har vi använt de rena diagnoserna (rensad lista) för att tagga upp nedanstående mappar för faktualitet. Maria har använt delar av detta för sitt första arbete med diagnos/finding.

Referenser

Skeppstedt M., H. Dalianis and G.H. Nilsson 2011. Retrieving disorders and findings: Results using SNOMED CT and NegEx adapted for Swedish, In the Proceedings of the LOUHI 2011, Third International Workshop on Health Document Text Mining and Information Analysis, Co-located with AIME 2011 Bled, Slovenia, July 6, 2011, CEUR-WS, volume 744, ISSN: 1613-0073, pp 11-17.

Stockholm EPR (Adverse Drug Event) ADE Corpus

Extrakt av journaler som har ADE ICD-10 koder: E27.3, G24.0, G25.1,
G44.4, G62.0, I42.7, I95.2, I42.7, L27.0, L27.1, N14.1, T78.3, T80.8, T88.7, T88.6

59766 tokens

Annoterare: Maria Kvist, sommaren 2016 (N.B. Aron Henriksson och Hercules Dalianis annoterade delmängder för att beräkna IAA)

Guidelines här:
http://dsv.su.se/polopoly_fs/1.243576.1439288669!/menu/standard/file/Guidelines_ADE.pdf

Named entities	**Types**	**Tokens**
Drug	853	1866
Disorder	976	3763
Finding	1533	3184
ADE Cue	130	341
Body Structure	297	1132
Overall	3789	10286

Attributes		
Negation	501	828
Speculation	333	467
Past	414	699
Future	250	354

Other	144	290
Overall	1642	2638

Relations

Indication	1156	1392
Adverse drug event	776	855
ADE Outcome	129	144
ADE Cause	205	228

Ligger på KEA-datorn: shared/annotations/ Stockholm EPR ADE Corpus

Referenser:

Henriksson, A., M. Kvist, H. Dalianis and M. Duneld. 2015. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. Journal of Biomedical Informatics, Vol 57: pp. 333-349, October 2015

Stockholm EPR Abbreviation Corpus

Annoterat förkortningar
Endast bedömningsfält

1) Medicinakuten:
6 batcher á 50 bedömningsfält = 300 bedömningsfält
batch5, file 1- 6, batch6, en fil
2 122 abbreviation tokens (335? unique), totalt 20 088 tokens I korpusen

2) Ortopedakuten:
2 batcher á 50 bedömningsfält = 100 bedömningsfält
tokens är inte räknade då detta material ej är använt till ngt ännu

Ligger på KEA under shared/annotations/ Stockholm EPR Abbreviation Corpus

Guidelines finns inte publicerade men jag har 2 A4 nedplitade på svenska med tankar.

Referenser

Isenius, N. 2012 Abbreviation detection in Swedish medical records. the development of SCAN, a Swedish clinical abbreviation normalizer. Master's thesis, Department of Computer and Systems Sciences, Stockholm University.

Isenius, N., S. Velupillai, and M. Kvist. Initial results in the development of SCAN. a Swedish clinical abbreviation normalizer. In CLEFeHealth 2012 workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis, Rome, 2012.

Stockholm EPR Detect-HAI Corpus

Annoterare: Elda Sparrelid och Mia Kvist. Våren 2012.

Klassificerade vårdtillfällen, patienter som har vårdrelaterade infektioner (VRI) och patienter som inte har. PPM Punkt prevalens mätningar från mars 2012, 120 patienter från 70 olika kliniker.

Totalt 213 vårdtillfällen, 129 vårdtillfällen klassificerade med VRI och 84 utan VRI

Totalt 1 549 908 tokens, varav 1 267 722 tokens med VRI och 282 197 tokens utan VRI.

Ligger på KEA under shared/annotations/Stockholm EPR Detect HAI Corpus

Referenser

Ehrentraut, C, H. Tanushi, H. Dalianis and J. Tiedemann. 2012. Detection of Hospital Acquired Infections in sparse and noisy Swedish patient records. A machine learning approach using Naïve Bayes, Support Vector Machines and C4.5. In the Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data, AND, December 9, 2012 held in conjunction with Coling 2012, Bombay.

Ehrentraut, C., Kvist, M., Sparrelid, E. and Dalianis, H. 2014. Detecting Healthcare-Associated Infections in Electronic Health Records - Evaluation of Machine Learning and Preprocessing Techniques, in the Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM 2014). Bodenreider, O., Oliveira, J.L., Rinaldi, F. (Eds.), Aveiro, Portugal,

Stockholm EPR Detect-HAI Corpus Reuma

Journaler från Reumatologiska kliniken. Totalt 67 stycken där 34 patienter var uppmärkta med VRI och 33 patienter utan VRI, med 292 vårdtillfällen längre än 48 timmar.

Ligger på KEA under shared/annotations/Stockholm EPR Detect HAI Corpus Reuma

Clinical Abbreviation Lists

1 229 expanderade kliniska förkortningar

Ligger på KEA under shared/annotations/Clinical Abbreviation Lists

Referenser

Kvist, M., and S: Velupillai 2014. SCAN: A Swedish Clinical Abbreviation Normalizer. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction* (pp. 62-73). Springer International Publishing.

Stockholm EPR Cervical Cancer Corpus

Annoterare: Karin Sundström och Mia Kvist. Två omgångar av annoteringar

Texter från journaler för patienter från onkologi och gynekologi med en cervixcancerdiagnos (ICD-10 kod C53). Texterna är annoterade för entiteterna Body Part, Finding och Disorder.

Body part: 2 103
Disorder: 1 059
Finding: 4 501

Ligger på KEA under shared/annotations/Stockholm EPR Cervical Cancer Corpus

Referenser

Weegar R., A. Pérez, A. Casillas and M. Oronoz (2018) Deep Medical Entity Recognition for Swedish and Spanish. International Workshop on Biomedical and Health Informatics in conjunction with 2018 IEEE International Conference on Bioinformatics and Biomedicin (BIBM).

Weegar R., M. Kvist, K. Sundström, S. Brunak, and H. Dalianis. 2015. Finding Cervical Cancer Symptoms in Swedish Clinical Text using a Machine Learning Approach and NegEx. In the proceedings of AMIA 2015 Annual Symposium, November 14-18, San Francisco, pp 1296-1305

Kliniskt vektorrum

vectors_1_10, skapat från en 1,2 Gb stor fil med klinisk text innehållandes 42 098 546 tokens som genererade 300 824 vektorer ligger i en 365 Mb stor fil. Kliniskt vektorrum 1_10, används i Rebecka Weegars studier, men också i Berg och Dalianis 2019.

Ligger på KEA under shared/annotations/Clinical vector space vectors_1_10

Referenser

Perez, A., R. Weegar, A. Casillas, K. Gojenola, M. Oronoz, M and H. Dalianis, H. 2017. Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora. *Journal of biomedical informatics*, *71*, 16-30.

Weegar, R. 2020. *Mining Clinical Text in Cancer Care* (Doctoral dissertation, Department of Computer and Systems Sciences, Stockholm University).

Berg, H. and H. Dalianis. 2019. Augmenting a De-identification System for Swedish Clinical Text Using Open Resources (and Deep learning). In the Proceedings of the Workshop on NLP and Pseudonymisation, in conjunction with the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), Turku, Finland, September 30, 2019

Stockholm EPR Structured and Unstructured ADE corpus, (SU-ADE Corpus)

Filen är på 24 Gb varav 18 Gb text och 6 Gb strukturerad information, 500 miljoner tokens från 1 314 646 patienter. Innehåller 5 typer av ADE ICD-10 koder och tillhörande texter

Ligger på KEA under shared/annotations/Stockholm_EPR_Structured_Unstructured_SU_ADE_Corpus/

Referenser

Bagattini, F., I. Karlsson, J. Rebane and P. Papapetrou. 2019. A classification framework for exploiting sparse multi-variate temporal features with application to ad- verse drug event detection in medical records. *BMC Medical Informatics and Decision Making*, 19(1):7, 12.

Bamba, M. and P. Papapetrou. 2019. Mining Adverse Drug Events Using Multiple Feature Hierarchies and Patient History Windows. In *Proceedings of the Workshop on Data Mining in Biomedical Informatics and Health- care DMBIH'19 in conjunction with IEEE International Conference on Data Mining, ICDM'19, Beijing*.

Etc, massor från Panos Papapetrous forskargrupp, Jonatha Rebane mm

Stockholm EPR Structured and Unstructured ADE Pseudo-corpus, (SU-ADE Pseudo-Corpus)

Ligger på KEA under
shared/annotations/Stockholm_EPR_Structured_Unstructured_SU_ADE_Pseudo_Corpus/

Samma som SU-ADE Corpus men pseudonymiserad

Stockholm EPR Gastro ICD-10 Corpus ver 2

4 985 patienter omfattande 6 062 epikriser med totalt 986 436 tokens från fyra gastrointestinala vårdenheter från åren 2007-2014.
Varje epikris är 163 tokens lång i genomsnitt, och är uppdelade i 10 ICD-10 kodblock från Kapitel 11 K-koderna innehållandes koderna K00-K93.
1,2 ICD-10 koder per epikris i snitt.
Antal unika koder 263

Dito

Stockholm EPR Gastro ICD-10 Pseudo Corpus ver 2

Samma som ovan men avidentifierad och pseudonymiserad

Ligger på KEA-datorn: shared/annotations/Stockholm EPR Gastro ICD 10 Corpus (ver2)

Referens
Remmer, Sonja. 2021. Automatic Diagnosis Code Assignment with KB-BERT ICD Classification – Using Swedish Discharge Summaries. Masteruppsats, Institutionen för data- och systemvetenskap, Stockholms universitet

Stockholm EPR Gastro ICD-10 Pseudo Corpus II

113 174 patienter omfattande 317 971 epikriser samt andra anteckningar med totalt 55 727 938 tokens från fyra gastrointestinala vårdenheter från åren 2007-2014.
Varje anteckning är 175,5 tokens lång i genomsnitt.
1,1 ICD-10 koder per anteckning i snitt.
Antal unika koder 415 (K-koder, Gastro)
Korpusen är avidentifierad och pseudonymiserad och kan delas med akademiska användare efter registrering och underskrift.

Ligger på KEA-datorn: shared/annotations/Stockholm EPR Gastro ICD-10 Pseudo Corpus II/

Referenser
Lamproudis, A., Olsen Svenning T., Torsvik T., Chomutare T., Budrionis A, Dinh Ngo P., Vakili T. and H. Dalianis. 2023. Using a Large Open Clinical Corpus for Improved ICD-10 Diagnosis Coding, to appear in the Proceedings of AMIA 2023, Annual Symposium, November 11-15. New Orleans, LA, USA.

Stockholm Clinical BERT text

17,8 Gb text, 2,8 x 10^9 ord i anteckningar från Health Bank 2007-2014 som används för att förträna Stockholm Clinical KB Bert

Ligger på KEA-datorn: shared/annotations/models/clinical_text_dump/data.txt

Pseudoversionen
Ligger på KEA-datorn: shared/annotations/models/clinical_text_dump/pseudo.txt

Språkmodeller: Clinical BERT

SweClin-BERT
Klinisk språkmodell baserad på Stockholm Clinical BERT text som är fortsatt förtränad på KB-BERT.

Ligger på KEA-datorn: shared/models/SweClin-BERT/

Referenser
Lamproudis, A., Henriksson, A., Dalianis, H. (2021). Developing a Clinical Language Model for Swedish: Continued Pretraining of Generic BERT with In-Domain Data. In Proc. of Recent Advances in Natural Language Processing (RANLP).

SweClin-BERT-replaced-vocabulary
Klinisk språkmodell baserad på Stockholm Clinical BERT text som är fortsatt förtränad på KB-BERT men där vokabulären är ersatt med en domän-specifik vokabulär. Två varianter finns baserat på olika sätt att initialisera vektorerna.

Ligger på KEA-datorn: shared/models/SweClin-BERT-replaced-vocabulary/

Referenser
Lamproudis, A., Henriksson, A. and H. Dalianis. 2022. Vocabulary Modifications for Domain-adaptive Pretraining of Clinical Language Models. In Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies – Volume 5: HEALTHINF, pp. 180-188.

SweClin-BERT-replaced-vocabulary-different-vocabulary-sizes
Klinisk språkmodell baserad på Stockholm Clinical BERT text som är fortsatt förtränad på KB-BERT men där vokabulären är ersatt med en domän-specifik vokabulär med olika storlekar (30k, 40k, 60k, 70k; 50k är SweClin-BERT-replaced-vocabulary). Det finns två varianter av varje baserat på olika sätt att initialisera vektorerna.

Ligger på KEA-datorn: shared/models/SweClin-BERT-different-vocabulary-sizes/

Referenser
Lamproudis, A. & Henriksson, A. (2023). On the Impact of the Vocabulary for Domain-Adaptive Pretraining of Clinical Language Models. In *International Joint Conference on Biomedical Engineering Systems and Technologies*, pp. 315-332. Cham: Springer Nature Switzerland.

Pure-SweClin-BERT
Klinisk språkmodell baserad på Stockholm Clinical BERT text som är förtränad från scratch. Den är

tränad i upp till 10 epoker (med checkpoints sparade efter varje epok); den bästa är efter 9 epoker och finns i en separat mapp: Best-Model-EPOCH-9).

Ligger på KEA-datorn: shared/models/Pure-SweClin-BERT/

Referenser
Lamproudis, A., Henriksson, A., Dalianis, H. (2022). Evaluating Pretraining Strategies for Clinical BERT Models. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC), pp. 410-416.

SweDeClin-BERT
Klinisk språkmodell baserad på en avidentifierad och pseudonymiserad Stockholm Clinical BERT text som är fortsatt förtränad på KB-BERT. Modellen är avidentifierad och pseudonymiserad och kan delas med akademiska användare efter registrering och underskrift.

Ligger på KEA-datorn: shared/models/SweDeClin-BERT/

Fine tunad för NER
Ligger på KEA-datorn: shared/models/SweDeClin-BERT-NER/

Referenser
Vakili, T., Lamproudis, A., Henriksson, A. and H. Dalianis. 2022. Downstream Task Performance of BERT Models Pre-Trained Using Automatically De-Identified Clinical Data, in Proceedings of the 13th International Conference on Language Resources and Evaluation, LREC 2022, Marseille, France, pp. 4245–4252.

Stockholm EPR ADE ICD-10 Corpus: “Sonjas ADE före 2007”
Ligger på KEA-datorn: shared/annotations/Stockholm EPR ADE ICD-10 Corpus/

Referenser
Lamproudis, A., Henriksson A. and H Dalianis. 2022. Evaluating Pretraining Strategies for Clinical BERT Models, in Proceedings of the 13th International Conference on Language Resources and Evaluation, LREC 2022, Marseille, France, pp.410–416

Patologirapporter
Cirka 16 487 prostata, 6 934 bröst och 8 740 cervixcancersvar samt datum, ICD koder och ATC läkemedelskoder för åren 2010-2021

Ligger på KEA-datorn: shared/Patologirapporter-2021