



Stockholm
University

Named entity recognition, rule based and machine learning based

Hercules Dalianis

Department of Computer and Systems Sciences (DSV)

hercules@csv.su.se



Overview

- Named entity recognition
- Rule based
 - Gazetteers
 - Regular expressions
- Machine learning based
 - Annotated data
 - Semi supervised methods – active learning
 - BERT Deep learning methods
- Named entity recognition and pseudonymisation
 - Privacy preserving machine learning



Stockholm
University

Named Entity Recognition

- Named Entity Recognition NER
- Named Entity Tagging
- Namnigenkänning (in Swedish)
- Named entities
 - In contrast to POS tagging
 - Noun, verbs, adjectives, etc...



MUC Message Understanding Conferences

- MUC-1 1987 – MUC-7 1997
- Challenges from DARPA (Defense Advanced Research Projects Agency in US.)
- *Named Entity Recognition (NER)* or Named Entity Tagging was first defined in the MUC 7 challenge as the identification of *personal names, locations, organisations* and *time points or dates*, also called TIMEX expressions in newswire text

Chinchor, N., & Robinson, P. (1997). MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding* (p. 29).



Stockholm
University

Named entities

- ENAMEX
 - Organization
 - Person
 - Location
- TIMEX
 - Date
 - Time
- NUMEX
 - Money
 - Percent

Named Entity Recognition

- Recognise
 - personal names, locations, organisations, products
time points, dates, measures, etc

<https://explosion.ai/demos/displacy-ent>



Examples on named entities

Morobito Consultants **ORG** said the company was engaged in **June of 2020 DATE** to prepare a “repair and restoration plan” for fixes needed under the state recertification requirements. At the time of the collapse **this week DATE**, the company said, roof repairs were underway but concrete restoration, which was to be handled by another firm, had not begun.

The collapse has stunned industry experts in the **Miami GPE** area, including **John Pistorino PERSON**, a consulting engineer who designed the **40-year DATE** reinspection program when he was consulting for the county in **the 1970s DATE**.



Stockholm
University

Extensions of NE

- Product
- Company

Coca Cola

Volvo



Stockholm
University

TIMEX

- Date
- Time
- Duration
- Frequency
- Event
- Absolute time/Relative time



Stockholm
University

Clinical entities

- Symptom
- Diagnosis
- Body part
- Drug
- Negated Symptoms and Diagnosis



Stockholm
University

Use of Named Entities

- News monitoring / Business Intelligence / Competitive intelligence
 - Find mentions a company names and products close to each other in a text
 - Opinion mining: Who is the opinion holder?
 - Automatic summarisation of text
 - Extract important passages
 - Categorisation or Classification of text
 - Feature engineering
 - **De-identification of a text**



Stockholm
University

Rule-based NER

- Gazetteers – name lists, personal names, locations, companies
 - First names, male and female
 - Last names
 - Streets, places, cities, countries, mountains, rivers.
 - Companies names



Rule-based NER (cont)

- Regular expressions
- Locations
 - Street, Avenue, Boulevard, St, Av, Bd,
 - Place, Square, Park, etc
 - <http://www.fallingrain.com/world/>
- Personal names
 - Mr, Mrs, Ms, Dr, von
 - Captain, Sir, Lady,

Rule-based NER (cont)

- Organizations
 - Ltd, Company, Factory, Holding, AB (Swedish)
- Timex
 - Seconds, minutes, hours, clock
 - day, months- July 10, Monday July 10,
 - am, pm
 - 2022-01-01, hours:min:sec min:sec



Stockholm
University

Swedish NER 2001 **Rules + Some training on corpus**

- Using 108 000 Swedish news articles
- Are matched possible names re-occurring twice then name. Should start with initial capital letter
- Check

Ett nytt rally kan pågå på E 4:an. Det kan också vara så att ett antal bilar rekognoseras inför en framtida rallytävling som uppges heta Cape to cape och har holländska arrangörer. Hittills har bilarna, som är registerade i Holland och är täckta av reklam, följt hastighetsbegränsningarna.

Deltagarna ska åka från Nordkap i **Norge** till Cape Terifa på Spaniens sydkust. Enligt tävlingens hemsida ska deltagarna starta den 27 september och målsättningen är att slå gällande rekord som, enligt hemsidan, finns registrerat i Guinness rekordbok. Något man knappast gör genom att följa hastighetsbegränsningarna.

- Vi vet inte alls syftet med det här, medger **Tor Petrell**, presstalesman vid polisen i **Stockholm**.

Ungefär vid tio tiden passerade knappt 15 bilar genom Stockholm.

- De körde helt lagenligt, säger **Tor Petrell** till svd.se.

Vid midnatt fick polisen i **Norrland** rapporter om att sju tävlingsbilar upptäckts norr om **Skellefteå**. I mörke kom bilarna till **Uppsala**.

- Vi tycker att de passerat spårlost förbi **Uppsala** län. De har åkt helt utan anmärkning, sade **Jan Sjökvist**, vakthavande befäl vid polisen i Uppsala, till svd.se strax före elva i förmiddags.

Flera polispatruller övervakade bilarnas färd längs E 4:an i **Uppsala** län. **Jan Sjökvist** uppges att Uppsala-polisen uppskattar antalet till mellan 5 och 15 och att de flesta är registerade i Holland.

- Det är inga extrema racerbilar. Men vi hoppas de har lämnat vårt län nu.

- **De passerade oss i Umeå** vid 4-tiden i mörke, säger **Kenneth Jonsson**, vakthavande befäl vid polisen i Umeå.

Polisen var efter rapporter om tävlingen förberedd och flera trafikpoliser var ute för att övervaka bilarnas framfart.

- Vi har inget att anmärka, säger Kenneth Jonsson.

Höll de hastighetsbegränsningarna?

- Ja, men de kom söderifrån i **fredags** och då gick det fort. Då trodde vi att målet var i Nordkap. Vi rapporterade ett par stycken för fortkörning. De var uppe i 130 på 70-väg. De borde vara precis söder om **Sundsvall** nu, om de följer hastighetsbegränsningarna. Om de står på är de väl i Stockholm, sade **Kenneth Jonsson** till svd.se strax före nio i mörke.

- **Vår gissning** är att de kommer att köra på E 4:an. De är nerlusade med reklam och vill väl synas.

Polisen i **Umeå** har förvarnat polisen söderut och enligt **Kenneth Jonsson** lär en ökad allmän övervakning ske längs E 4:an i dag. På valborgsmässoafton i år körde ett hundratal lyxbilar rally mellan **Stockholm** och **Malmö**. Den tävlingen hette Gumball, flera av bilarna stoppades för fortkörning av polisen. Hastigheter på över 200 kilometer i timmen uppmättes i vissa fall.

[Jonas Elgh](#)

Found names: 1 **Tor Petrell**; 1 **Kenneth Jonssons**; 1 **Sjökvist**; 3 **Kenneth Jonsson**; 1 **Tor Petrells**; 1 [Jonas Elgh](#)

Found locations: 4 **Uppsala**; 3 **Umeå**; 1 **Malmö**; 1 **Sundsvall**; 1 **Norge**; 1 **Norrland**; 3 **Stockholm**; 1 **Skellefteå**

Found companies:

Found times: 1 **fredags**; 2 **Tor**; 2 **Jan**;



Stockholm
University

Dalianis, H., & Åström, E. (2001). SweNam-A Swedish Named Entity recognizer. Its construction, training and evaluation. *Technical Report TRITA-NA-PO113*.

92 percent precision and 46 percent recall of the named entities of a text from rule based recognition with training.



Stockholm
University

Swedish NER for clinical text 2009

- Manual annotation of clinical texts
- Three annotators, annotated the same texts
- PHI-Protected Health Information
- PII-Protected Personally Identifiable Information



28 PHI-classes (HIPAA) - American standard

- Account_Number, Age, Age_Over_89, Biometric_Identifier, Date_Part, Full_Date, Year, First_Name, Last_Name, Patient_First_Name, Patient_Last_Name, Relative_First_Name, Relative_Last_Name, Clinician_First_Name, Clinician_Last_Name, Location, Country, Municipality, Organization, Street_Address, Town, Health_Care_Unit, Device_Identifier_and_Serial_Number, Ethnicity, Fax_Number, Phone_Number, Relation, Uncertain



Example on annotations

Planeringsansvarig SSK Tjänstgörande

Patientansvarig läkare

 <First_Name>Mohamed</First_Name>
 <Last_Name>Åström</Last_Name>
Kontaktorsak Ramlat i hemmet <Full_Date>10/5-
29</Full_Date> och krampat <Date_Part>6/5</Date_Part>.
Hade inte ätit eller druckit på 4 dygn.

Hälsohistoria/ vårderf. Se läkare anteckningar.

Närstående Dotter <First_Name>Jessica</First_Name>
<Last_Name>Fredriksson</Last_Name> tel:
<Phone_Number>071-54639208</Phone_Number> , tel hem
91 38 45<Last_Name>Fredriksson</Last_Name> tel.

<Phone_Number>0392-857461</Phone_Number>

Social bakgrund Bor på gruppboende,
<Health_Care_Unit>KS</Health_Care_Unit> i
<Location>Ladugårdsgärdet</Location>.

Original annotation on Stockholm EPR PHI Corpus



Stockholm
University

Exact matches		Relevant	Retrieved	Corpus	Precision	Recall	F-score
AGE	Class						
	Account_Number	0	0	1	NaN	0.000000	NaN
	Age	40	45	54	0.888889	0.740741	0.808081
DATE	Age_Over_89	0	0	3	NaN	0.000000	NaN
	Biometric_Identifier	0	0	4	NaN	0.000000	NaN
	Date_Part	629	682	843	0.922287	0.746145	0.824918
PERSON-NAME	Full_Date	338	427	503	0.791569	0.671968	0.726882
	Year	15	22	62	0.681818	0.241935	0.357143
	First_Name	0	0	13	NaN	0.000000	NaN
LOCATION	Last_Name	0	0	5	NaN	0.000000	NaN
	Patient_First_Name	6	14	75	0.428571	0.080000	0.134831
	Patient_Last_Name	0	0	3	NaN	0.000000	NaN
PHONE-NUMBER	Relative_First_Name	67	72	128	0.930556	0.523438	0.670000
	Relative_Last_Name	8	8	23	1.000000	0.347826	0.516129
	Clinician_First_Name	541	612	735	0.883987	0.736054	0.803267
PHONE-NUMBER	Clinician_Last_Name	706	764	901	0.924084	0.783574	0.848048
	Location	0	0	3	NaN	0.000000	NaN
	Country	5	7	27	0.714286	0.185185	0.294118
PHONE-NUMBER	Municipality	3	14	34	0.214286	0.088235	0.125000
	Organization	0	1	60	0.000000	0.000000	NaN
	Street_Address	0	0	12	NaN	0.000000	NaN
PHONE-NUMBER	Town	7	12	52	0.583333	0.134615	0.218750
	Health_Care_Unit	910	1 162	1 747	0.783133	0.520893	0.625645
	Device_Identifier_and_Serial_Number	0	0	6	NaN	0.000000	NaN
PHONE-NUMBER	Ethnicity	0	0	9	NaN	0.000000	NaN
	Fax_Number	0	0	5	NaN	0.000000	NaN
	Phone_Number	56	65	130	0.861538	0.430769	0.574359
PHONE-NUMBER	Relation	458	473	714	0.968288	0.641457	0.771693
	Uncertain	0	1	18	0.000000	0.000000	NaN
	Total	3789	4 381	6 170	0.864871	0.614100	0.718226



Stockholm
University

Consensus eight annotation classes

- *Age*
- *Date_Part*
- *Full_Date*
- *First_Name*
- *Last_Name,*
- *Health_Care_Unit*
- *Location*
- *Phone_Number*



Annotation classes

• Age	56
• Full date	710
• Date part	500
• First name	923
• Last name	928
• Location	1,021
• Health care unit	148
• Phone number	135
Sum:	4,421



Machine learning based NER using CRF

- Hidden Markov Model (HMM) is a generative model
 - to model a chain of possibilities of signal, molecular sequences, words, etc
- CRF is a discriminative model slightly resembling HMM, but better adapted for NER

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.. Proceedings of the 18th International Conference on Machine Learning. 282-289. 2001.



Stockholm
University

Other approaches with ML for NER

- Support Vector Machines
- Random Forest
 - Can also be used for NER but not as good as CRF



Stockholm
University

CRF implementations

- Stanford NER
- CRF++ (Yet Another CRF toolkit)
- CRFsuite
- Mallet CRF
 - And many others



Stockholm
University

Example on annotations

For training of CRF or fine tuning of BERT models



Stockholm
University

SGML format

<PER>Alex</PER> is going with <PER>Marty A.
Rick</PER> to <LOC>Los Angeles</LOC> .

Conll IOB format (short for inside, outside, beginning) (BIO-format)

Alex	S-PER
is	O
going	O
with	O
Marty	B-PER
A.	I-PER
Rick	E-PER
to	O
Los	B-LOC
Angeles	E-LOC
.	O



Raw text i SGML format

The Groningen Meaning Bank

<https://gmb.let.rug.nl/data.php>

Doctors say Israeli Prime Minister Ariel Sharon remains in critical but stable condition - 10 days after suffering a massive stroke.

Jerusalem's Hadassah hospital issued the update on Mr. Sharon's condition Saturday evening after the end of the Jewish sabbath.

The hospital said medical tests showed that Mr. Sharon has activity in both of his brain lobes.

For several days, doctors have been reducing the level of sedatives, with the intent of drawing the 77-year-old Mr. Sharon out of an induced coma. However, he has not shown any signs of waking up.

Conll format

Doctors	NN	N[nom]	doctor	O
say	VBP	S[dcl]\NP/S[dcl]	say	O
Israeli	JJ	N[nam]/N[nam]	israeli	O
Prime	NNP	N[nam]/N[nam]	prime	O
Minister	NNP	N[nam]/N[nam]	minister	O
Ariel	NNP	N[nam]/N[nam]	ariel	I-PER
Sharon	NNP	N[nam]	sharon	I-PER
remains	VBZ	S[dcl]\NP	remain	O
in	IN	S[dcl]\NP\S[dcl]\NP/NP	in	O
critical	JJ	N[nom]/N[nom]	critical	O
.....				
.....				
Jerusalem	NNP	N[nam]	jerusalem	I-LOC
's	POS	NP/N[nom]\NP	's	O
Hadassah	NNP	N[nom]/N[nom]	hadassah	I-LOC
hospital	NN	N[nom]	hospital	O
issued	VBD	S[dcl]\NP/NP	issue	O
the	DT	NP/N[nom]	the	O
update	NN	N[nom]	update	O
on	IN	NP\NP/NP	on	O
Mr.	NNP	N[nam]/N[nam]	mr.	I-LOC
Sharon	NNP	N[nam]	sharon	I-PER
's	POS	NP/N[nom]\NP	's	O
condition	NN	N[nom] condition	O	
Saturday	NNP	S[dcl]\NP\S[dcl]\NP/S[dcl]\NP\S[dcl]\NP	saturday	I-DAT
evening	NN	S[dcl]\NP\S[dcl]\NP	evening	I-TIM
after	IN	S[dcl]\NP\S[dcl]\NP/NP	after	I-TIM



Stockholm
University

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

Figure 8.7 NER as a sequence model, showing IO, BIO, and BIOES taggings.

We've also shown two variant tagging schemes: IO tagging, which loses some information by eliminating the B tag, and BIOES tagging, which adds an end tag E for the end of a span, and a span tag S for a span consisting of only one word. A sequence labeler (HMM, CRF, RNN, Transformer, etc.) is trained to label each token in a text with tags that indicate the presence (or absence) of particular kinds of named entities.

Words	POS	Short shape	Gazetteer	BIO Label
Jane	NNP	Xx	0	B-PER
Villanueva	NNP	Xx	1	I-PER
of	IN	x	0	O
United	NNP	Xx	0	B-ORG
Airlines	NNP	Xx	0	I-ORG
Holding	NNP	Xx	0	I-ORG
discussed	VBD	x	0	O
the	DT	x	0	O
Chicago	NNP	Xx	1	B-LOC
route	NN	x	0	O
.	.	.	0	O

Figure 8.16 Some NER features for a sample sentence, assuming that Chicago and Villanueva are listed as locations in a gazetteer. We assume features only take on the values 0 or 1, so the first POS feature, for example, would be represented as $\mathbb{1}\{\text{POS} = \text{NNP}\}$.



Stockholm
University

Our created annotated data set

- Stockholm EPR PHI Corpus



Stockholm
University

Consensus eight annotation classes

- *Age*
- *Date_Part*
- *Full_Date*
- *First_Name*
- *Last_Name,*
- *Health_Care_Unit*
- *Location*
- *Phone_Number*



Annotation classes

• Age	56
• Full date	710
• Date part	500
• First name	923
• Last name	928
• Location	1,021
• Health care unit	148
• Phone number	135
Sum:	4,421



Eight annotation classes training and test using Stanford NER (CRF)

Class	Annotated	Retrieved	Relevant	Exact matches			Partial matches		
				Precision	Recall	F-score	Precision	Recall	F-score
Age	56	45	37	0.822222	0.660714	0.732673	0.904762	0.778061	0.836642
Date_Part	710	654	617	0.943425	0.869014	0.904692	0.946196	0.871730	0.907438
Full_Date	500	426	342	0.802817	0.684000	0.738661	0.931665	0.802106	0.862045
First_Name	923	749	713	0.951936	0.772481	0.852871	0.954606	0.773772	0.854729
Last_Name	928	816	777	0.952206	0.837284	0.891055	0.961653	0.845484	0.899835
Health_Care_Unit	1021	689	559	0.811321	0.547502	0.653801	0.921497	0.608116	0.732705
Location	148	73	54	0.739726	0.364865	0.488688	0.778539	0.379129	0.509933
Phone_Number	135	86	80	0.930233	0.592593	0.723982	0.954195	0.613105	0.746535
Total	4421	3538	3179	0.898530	0.719068	0.798844	0.941190	0.751441	0.835680

Additional file 5 (Table S5) - Results of the manual Consensus Gold standard using ten-fold cross-evaluation



Features

- Initial capital letter - `string.istitle()`
- Lower – `string.islower()`
- Upper - `string.isupper()`
- Alpha – `string.alpha(string)`
- Numeric - `string.isnumeric()`
- Length of word/string - `len(string)`



Stockholm
University

HB Deid

- CRF Suite
- Dictionaries (Gazetteer)
- Regular expressions
 - Phone numbers
 - Social Security Numbers (Personnummer)
 - Hybrid approach

Stockholm EPR PHI Corpus using CRF Suite



	EPR PHI			EPR PHI + Domain			EPR PHI + SUC		
	P %	R %	F ₁	P %	R %	F ₁	P %	R %	F ₁
First Name	95.05	92.78	0.94	95.50	91.41	0.93	94.51	92.24	0.93
Last Name	97.02	92.20	0.94	96.39	90.36	0.93	96.93	0.93	0.95
Phone Number	92.81	81.52	0.87	94.58	84.32	0.89	96.14	72.14	0.82
Age	79.29	60.95	0.68	85.09	71.27	0.77	89.67	76.21	0.82
Full Date	98.62	99.15	0.99	96.34	94.74	0.96	95.82	93.28	0.94
Date Part	97.06	95.68	0.96	97.45	94.73	0.96	95.46	90.08	0.93
Health Care Unit	86.11	66.40	0.75	88.62	72.79	0.80	85.45	67.38	0.75
Location	74.07	73.70	0.72	76.05	59.89	0.66	62.40	70.92	0.65
Overall	93.76	86.53	0.90	94.66	86.72	0.91	92.31	84.97	0.88

Table 4: Entity-based evaluation for CRF with tenfold cross-validation. The mean is presented for each label. The highest F₁ scores are highlighted for each class.

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

Semisupervised learning

Active learning

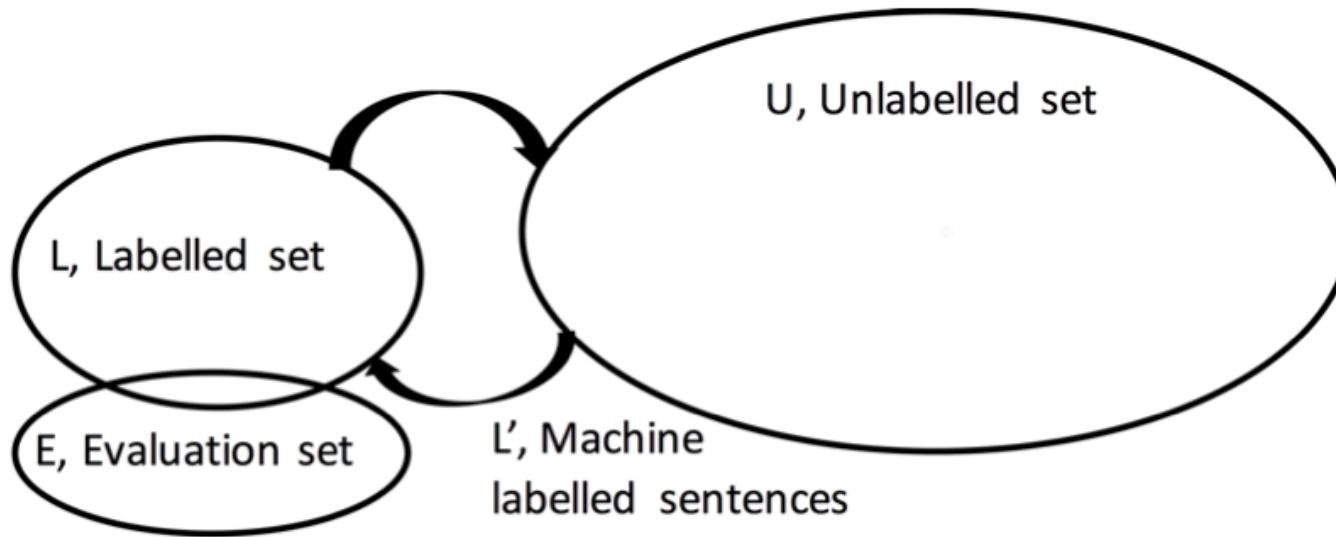


Figure 1: Figure over the three different data sets: The *labelled set L*, the *unlabelled set U* and the *labelled evaluation set E*.

Increase annotated training data

Used sets	L	S	NS
Age	28	565	613
Date Part	328	5,984	6,725
First Name	589	7,779	10,416
Full Date	297	4,140	4,404
Health Care Unit	771	6,213	9,035
Last Name	627	7,995	11,184
Location	51	1,642	2,257
Phone Number	104	243	461
Overall	2,795	34,561	40,695

Table 5: The table presents the amount of annotated tokens in the original dataset (L), as well as the amount in the automatically labelled dataset with thresholds (S) and without thresholds.

- Automatic annotated data 20 times larger than the manually annotated
- The selection is carried out on marginal scores, i.e. how likely they are correct.



Precision stable, Recall Increases

	Unigram Tagger	Baseline trained on labelled data L	Non-selection	Selection
Precision	92.82%	95.80%	94.14%	94.20%
Recall	44.55%	84.93%	88.16%	89.20%
F1	76.29	90.04	91.05	91.63
Leakage	33.90%	12.84%	9.40%	8.16%

Table 1: Comparison of precision, recall, F1 and leakage for all four models. The Unigram Tagger is described in subsection 3.3. Training Baseline is the model with only manually labelled data from L , with results further presented in Table 2. Selection is the semi-supervised model described in subsection 3.4, with results further presented in Table 3. Non-selection (NS) is the model where all automatically data is included and is further described in Table 4.

- NLTK Unigram tagger, tag word with its label-not seen word is assumed being PHI/PII

Berg, H. and H. Dalianis. 2020. A Semi-supervised Approach for De-identification of Swedish Clinical Text. Proceedings of 12th Conference on Language Resources and Evaluation, LREC 2020, May 13-15, Marseille, pp. 4444-4450,



Stockholm
University

KB BERT + Clinical Health Bank BERT

- KB BERT 18 Gb general Swedish from National Library of Sweden (Kungliga Biblioteket)
 - Swedish news text
 - Swedish Wikipedia
 - Swedish governmental reports
- Health Bank 17.8 Gb of clinical text

Continued pretraining

- Continued pretraining from KB BERT on Health Bank creating KB-Clinical BERT



Evaluation on Stockholm EPR PHI

Corpus

CRFsuite incl
RegExp

KB+Clinical Health
Bank BERT

	P %	R %	F ₁	Class (instances)	KB-BERT	M-BERT
First Name	95.05	92.78	0.94	First Name (195)	0.9385	0.9077
Last Name	97.02	92.20	0.94	Last Name (213)	0.9531	0.9296
Phone Number	92.81	81.52	0.87	Phone Number (21)	0.9048	0.8571
Age	79.29	60.95	0.68	Age (9)	1.0000	0.7778
Full Date	98.62	99.15	0.99	Full Date (83)	0.9518	0.9518
Date Part	97.06	95.68	0.96	Date Part (131)	0.9847	0.9824
Health Care Unit	86.11	66.40	0.75	Health Care Unit (293)	0.8737	0.8396
Location	74.07	73.70	0.72	Location (19)	0.7895	0.4221
Overall	93.76	86.53	0.90	Organisation (10)	0.5000	0.5000
				Weighted average	0.9220	0.8899



Combine CRF and Regular expressions and list

- Regular expressions works best on structured info such as timepoints, dates and social security number (personnummer)
- Machine learning based - works best on unstructured but need training instances i.e. annotations
- Combine both approaches to a hybrid approach.



Stockholm
University

Privacy preserving

- Privacy preserving of sensitive text for research and for following GDPR
 - Clinical text
 - Customer services
 - Questionaries
 - Lawsuits
- What to remove and how much?
- PHI Protected Health Information



Stockholm
University

Named entity recognition and pseudonymisation

- HB Deid, <https://hbdeid.dsv.su.se>

De-identification of free text by Named Entity Recognition (NER)

Two steps

- First step NER (Identify)
- Second step
 - Remove
 - Mask
 - Mark with identifier/class name e.g. First Name
 - Replace with surrogates or pseudonyms

De-identification and pseudonymisation

- Identification:
 - Hans fru Anna (His wife Anna) =>
 - Hans fru <First_Name>Anna</First_name>
- De-identification/Anonymisation: Klass
 - Hans fru Anna (His wife Anna) =>
 - Hans fru <First_Name>
- With pseudonymisation:
 - Hans fru Anna (His wife Anna) =>
 - Hans fru **Lena**
- With masking:
 - Hans fru Anna (His wife Anna) =>
 - Hans fru **XXX**



Pseudonymisation

Resynthesis or “re-identification with fake data, surrogates or pseudonyms

- Name change (to any name)
- Keep the gender of the patient.
- Names for both gender, tex Kim, Pat, Robin.
- Names on the genitive form (genitive s).
- Repetition of the same fictive name
 - Keep in the family
- Unknown and Misspelled names
 - Spelling corrections



Dictionaries

- While common names are replaced with other common names, uncommon names are replaced with uncommon names.
- The uncommon names in the dictionaries are 123,000 female first names, 121,000 male first names and 35,000 last names.
- The web interface for the HB-demo is written in Flask that in turn is coded in Python.



Stockholm
University

HIPS

- Hiding In Plain Sight (HIPS)
- Remove tags for not revealing the missed ones



Bootstrapping models for privacy

- The data used to train HB Deid Demo is first trained directly on the original set of annotated sensitive electronic patient records in Swedish - the Stockholm EPR PHI Corpus.
- Then the created model is used to machine annotate the freely available Webbnyheter 2012, (at Språkbanken), that is semi-manually annotated for the NER classes PER and LOC and ORG and MISC.
- Then the double machine annotated Webbnyheter 2012 is used as a privacy protected model for HB Deid.



HB Deid - Automatisk aidentifiering av text

Den här versionen av HB Deid är tränad på icke-känsliga annoterade tidningstexter (Webbnyheter 2012), snabbannoterad genom bootstrapping med en modell tränad på Stockholm EPR PHI Corpus. Den kan användas för att identifiera och byta ut personnamn, platser, telefonnummer, åldrar, personnummer och telefonnummer.

Inga av texterna som matas in av användaren sparas.

- *Ersättare*: Annat ord som val för innebär att texten pseudonymiseras
- *Viktad*: Ökande vikt ger högre täckning, men också sämre precision
- *+ Regler*: Ja innebär att regler används för att identifiera bland annat telefonnummer och personnummer

Ersättare

Annat ord

Viktad

Ingen

+ Regler

Ja

Din text

(Den här texten är fiktiv)

2020-07-27

22-årig patient med TBC inkom idag till Lungkliniken, Karolinska Universitetssjukhuset, Solna. Patienten har varit i Somalia. Han åtföljdes av sin mor Samira Ahmed, tel 070-578 38 32, Rådmansgatan 4, 114 24 Stockholm. Samira var mycket orolig för sin son. Patienten kom tillbaka till Sverige den 20 juli från Somalia. Patienten har personnummer 19981212-1237. I Mogadishu badade patienten i havet och blev biten av stingrocka som orsakade ett sår över vaden. Patienten har feber som har pendlat mellan 38 och 40 grader i två veckor och han är medtagen. Har ordinerat Flukloxacillin, 5 mg dagligen.

2020-07-28

Ingen feber Såret på vaden läker bra.

2020-07-30

Vilande TBC, ska kontrolleras regelbundet på Lungkliniken. Patienten skrivs ut.



(Den här texten är fiktiv)

2020-07-27

22-årig patient med TBC inkom idag till Lungkliniken, Karolinska Universitetssjukhuset, Solna. Patienten har varit i Somalia. Han åtföljdes av sin mor Samira Ahmed, tel 070-578 38 32, Rådmansgatan 4, 114 24 Stockholm. Samira var mycket orolig för sin son. Patienten kom tillbaka till Sverige den 20 juli från Somalia. Patienten har personnummer 19981212-1237. I Mogadishu badade patienten i havet och blev biten av stingrocka som orsakade ett sår över vaden. Patienten har feber som har pendlat mellan 38 och 40 grader i två veckor och han är medtagen. Har ordinerat Flukloxacillin, 5 mg dagligen.

2020-07-28

Ingen feber. Såret på vaden läker bra.

2020-07-30

Vilande TBC, ska kontrolleras regelbundet på Lungkliniken. Patienten skrivas ut.

(Den här texten är fiktiv)

2020-06-01

22-årig patient med TBC inkom idag till en vårdenhets, Akademiska Sjukhuset Universitetssjukhuset, Handen. Patienten har varit i Nigeria. Han åtföljdes av sin mor Stensdtr Lindgren, tel 070-626 66 66, Vårby Gård 4, 114 24 Stockholm. Stensdtr var mycket orolig för sin son. Patienten kom tillbaka till Sverige den 20-07 från Nigeria. Patienten har personnummer 191212-1212. I Shinyanga badade patienten i havet och blev biten av stingrocka som orsakade ett sår över vaden. Patienten har feber som har pendlat mellan 38 och 40 grader i två veckor och han är medtagen. Har ordinerat Flukloxacillin, 5 mg dagligen.

2020-06-02

Ingen feber. Såret på vaden läker bra.

2020-06-04

Vilande TBC, ska kontrolleras regelbundet på en vårdenhets Patienten skrivas ut.



Stockholm
University





Chat GPT -

<https://chat.openai.com/chat>

Skriv ut alla namn, platser, adresser i denna text:

(Print all names, places and addresses in this text:)

"2020-07-27

22-årig patient med TBC inkom idag till Lungkliniken, Karolinska Universitetssjukhuset, Solna. Patienten har varit i Somalia. Han åtföljdes av sin mor Samira Ahmed, tel 070-578 38 32, Rådmansgatan 4, 114 24 Stockholm. Samira var mycket orolig för sin son. Patienten kom tillbaka till Sverige den 20 juli från Somalia. Patienten har personnummer 19981212-1237. I Mogadishu badade patienten i havet och blev biten av stingrocka som orsakade ett sår över vaden. Patienten har feber som har pendlat mellan 38 och 40 grader i två veckor och han är medtagen. Har ordinerat Flukloxacillin, 5 mg dagligen.

2020-07-28

Ingen feber. Såret på vaden läker bra.

2020-07-30

Vilande TBC, ska kontrolleras regelbundet på Lungkliniken. Patienten skrivs ut.



Stockholm
University

Answer from Chat GPT

- Namn: 22-årig patient
Plats: Lungkliniken, Karolinska Universitetssjukhuset, Solna
Adress: -

Namn: Samira Ahmed
Plats: Stockholm
Adress: Rådmansgatan 4, 114 24 Stockholm
Telefonnummer: 070-578 38 32

Namn: -
Plats: Somalia
Adress: -

Namn: -
Plats: Mogadishu
Adress: -

Sjukdom: TBC
Medicin: Flukloxacillin, 5 mg dagligen
Personnummer: 19981212-1237

Ta bort alla namn, platser, adresser i denna text:

(Remove all names, places and addresses in this text:)

Answer from Chat GPT

- 2020-07-27 22-årig patient med TBC inkom idag till sjukhus. Patienten har varit i ett land utanför Sverige. Han åtföljdes av en person. Personen var orolig för patienten. Patienten kom tillbaka till Sverige från det andra landet. Patienten har haft feber i två veckor och är medtagen. Läkaren har ordinerat en medicin.
- 2020-07-28 Ingen feber, såret läker bra.
- 2020-07-30 Vilande TBC, ska kontrolleras regelbundet på sjukhuset. Patienten skrivas ut.



Remove all names, places and addresses in this text:)

Answer from Chat GPT Translated to English

- 2020-07-27

22-year-old patient with TB was admitted to hospital today. The patient has been to **a country** outside Sweden. He was accompanied by **one person**. **The person** was worried about the patient. The patient came back to Sweden from the **other country**. The patient has had a fever for two weeks and is admitted. The doctor has prescribed **a medicine**.
- 2020-07-28 No fever, the wound is healing well.
- 2020-07-30 Dormant TB, must be checked regularly at the hospital. The patient is discharged.



Stockholm
University

Ethical problems to use Chat GPT

- Ethical problems to use Chat GPT
 - For de-identification
 - For clinical text mining
 - Etc.



Stockholm
University

Privacy preserving clinical BERT model

- Work by Thomas Vakili
- How much will de-identification and pseudonymisation of the Health Bank text effect downstream tasks?



Health Bank (Text) de-identified

- Health Bank 17.9 Gb text de-identified with KB+Clinical Health Bank BERT. It took 10 days!
- 3 227 572 231 tokens => 1000 bibles => 1 m³

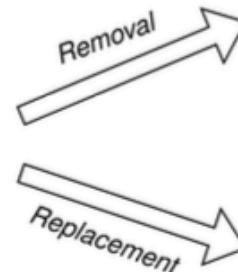
Class	PHIs	Per cent
First_Name	12 525 688	0,39%
Last_Name	14 332 309	0,44%
Location	3 158 031	0,10%
Phone_Number	1 262 313	0,04%
Age	2 064 111	0,06%
Date_Part	19 374 711	0,60%
Full_Date	10 459 935	0,32%
Health_Care_Unit	19 659 127	0,61%
Organisation	1 078 115	0,03%
Summa	83 914 340	2,60%



PHI Type	# Predicted Instances	NER Recall	NER Precision
<i>Health Care Unit</i>	19,659,127	80%	87%
<i>Partial Date</i>	19,374,711	83%	94%
<i>Last Name</i>	14,332,309	97%	96%
<i>First Name</i>	12,525,688	97%	98%
<i>Full Date</i>	10,459,935	55%	77%
<i>Location</i>	3,158,031	89%	85%
<i>Age</i>	2,064,111	35%	47%
<i>Organisation</i>	1,078,115	36%	71%
<i>Phone Number</i>	1,262,313	40%	63%

Table 1: The PHI types in order of frequency as classified by the de-identification system. The per-class recalls and precisions for the NER model are also displayed and were calculated on the test data from [Velupillai et al. \(2009\)](#). In total, 83,914,340 sensitive entities are found in 49,715,558 sentences.

Pat arrives to hospital with broken tibia.
Anaesthetic given by nurse **Andersson**.
Sent home with cast on leg.



Pat arrives to hospital with broken tibia.

Sent home with cast on leg.

Pat arrives to hospital with broken tibia.
Anaesthetic given by nurse **Jonsson**.
Sent home with cast on leg.

KB-BERT + Pseudo The data used to train this model has had all sensitive entities (as listed in Table 1) replaced with realistic surrogates of the same entity class.

KB-BERT + Filtered This model is built using the dataset where all sentences found to contain sensitive data have been *removed*. This filtered version of the dataset is 19% smaller than the version used to train *KB-BERT + Pseudo*.



Model	ICD-10	PHI	Clinical Entity	Factuality	Factuality	ADE
	Classification	NER	NER	Classification	NER	Classification
KB-BERT	0.799	0.91	0.803	0.635	0.630	0.183
KB-BERT + Real	0.833	0.941	0.858	0.732	0.682	0.199
KB-BERT + Filtered	0.833	0.929	0.854	0.731	0.672	0.199
KB-BERT + Pseudo	0.832	0.941	0.861	0.736	0.684	0.191

Table 3: The table compares the downstream performances of each BERT model. *KB-BERT* and *KB-BERT + Real* are used as baselines. *KB-BERT* is also the starting point for the continued pre-training of all three models, as described in Section 4.2. All values are F₁-scores and the best results are bolded.

SweDeClin-BERT

- Swedish De-identified language model based on KB-BERT and 2 million de-identified and pseudonymised patient records
- 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,

Privacy risks with de-identified BERT

- The performance of the de-identification system suggests that the overwhelming majority of sensitive data are removed from the training data of our models. If only 3% of all names in the data used for domain adaptation are sensitive, and the risk of exposing *any* name is less than 10% (Jagannatha et al., 2021), then the risk of exposing a *real* name is very small.

Jagannatha, A., Rawat, B. P. S., and Yu, H. (2021). Membership Inference Attack Susceptibility of Clinical Language Models. *arXiv:2104.08305 [cs]*, April. arXiv: 2104.08305.

Hacking BERT by generating text

- MIMIC III database
- Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In 30th USENIX Security Symposium (USENIX Security 21) (pp. 2633-2650).
- Lehman, E., Jain, S., Pichotta, K., Goldberg, Y., & Wallace, B. C. (2021, April). Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?. In North American Chapter of the Association for Computational Linguistics (NAACL).
- Vakili, T. and H. Dalianis. 2021. Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations. In the Proceedings of the Association for the Advancement of Artificial Intelligence AAAI Fall 2021 Symposium in HUman partnership with Medical Artificial iNtelligence (HUMAN.AI), November 4-6, 2021.

Clinical Note

mr coleman is [...] described a single jerk lasting [...] frequency of the jerks [...]



Generated Text

[...] \" eeg now with coleman ? \"
spur artifact [...] myclonic jerking
postures [...]



	First name	Last name	Name + Condition	Name + Wrong condition
Lehman et al. (2021)	0.94%	3.14%	23.53%	28.33%
$k = 1000$	1.04%	3.61%	24.06%	28.28%
$p = 0.99$	1.28%	3.76%	24.72%	28.25%
$p = 0.95$	1.10%	3.81%	25.51%	29.33%

- MIMIC clinical data
 - 3% identified where only 25% is identified with a condition some risk for identification



Stockholm
University

KB Bert + Deidentified and pseudonymised Health Bank => SweDeClin-BERT

- A mix of non-sensitive and sensitive data that has been de-identified
- 3% identified where only 25% is identified with a condition
- 14 million last names \times 3 % = 420 000 last names not de-identified => 25 % 100 000 last names and conditions but a mix of names from KB BERT and Health Bank
- Still a risk but much much smaller.

Clinical (Named) Entitites

- An other type of named entity recognition



Manually annotation for detection of clinical named entities

- Program modules for detection of
 - Symptom and diagnosis
 - Negation
 - Uncertainty
 - Period of time

76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan.
Inkommer med centrala bröstmärter utan utstrålning.

- 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central **chest pain** without **radiation**.



Manually annotation for detection of clinical named entities

- Program modules for detection of
 - Symptom and diagnosis
 - Negation
 - Uncertainty
 - Period of time

76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan.
Inkommer med centrala bröstmärter utan utstrålning.

- 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central **chest pain without radiation**.



Manually annotation for detection of clinical named entities

- Program modules for detection of
 - Symptom and diagnosis
 - Negation
 - Uncertainty
 - Period of time

76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan.
Inkommer med centrala bröstmärter utan utstrålning.

- 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.



Manually annotation for detection of clinical named entities

- Program modules for detection of
 - Symptom and diagnosis
 - Negation
 - Uncertainty
 - Period of time

76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan.
Inkommer med centrala bröstmärter utan utstrålning.

- 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.



Manually annotation for detection of clinical named entities

- Program modules for detection of
 - Symptom and diagnosis
 - Negation
 - Uncertainty
 - Period of time

76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan.
Inkommer med centrala bröstmärter utan utstrålning.

- 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.



Manually annotation of internal medicine emergency unit patient records

- Over 8,000 annotated entities
 - 1,998 Disorders
 - 3,822 Findings
 - 1,539 Drugs
 - 750 Body parts

Pre-annotation trained on internal medicine emergency unit records, applied on obstetrics/gynecology and oncology Annotation and correction by manual annotators



Stockholm
University

1	Remiss från Akuten, [REDACTED] under diagnos urtikaria i samband med Ipron. Referral from Emergency unit with diagnosis urtikaria in connection with Ipron	Disorder	Adverse_Drug_Reaction	Drug
2	Frisk.	Finding		
3	Reumatism sedan småbarnsåldern.	Disorder	since childhood	Indication
4	Haft någon reumatolog i [REDACTED] men aldrig sökt i Sverige.			Indication
5	Använder bara Brufen vid behov. uses Brufen when needed		Drug	
6	Hittills har det gått bra uteslutande besvär i tårna, inga andra leder.	Body_Part	Body_Part	
7	I samband med tandläkaringrepp när han tog bort en visdomstand och en vanlig tand som sprack ganska mycket lokalbedövning given, så åkte han hem och tog Ipron pga ganska mycket In connection with dental procedures when he had removed a wisdom tooth and regular tooth that cracked plenty of local anesthetic given, he went home and took an Ipron	Body_Part	Body_Part	Drug
8	värk, en halvtimme senare kläda, i munnen och sedan kraftiga utslag över hela kroppen utan andningspåverkan tillkallade ambulans. due to lot of pain, half an hour later itching in mouth and then severe rash over the whole body no respiratory distress called ambulance	Finding	Finding	Adverse_Drug_Reaction
9	Fick 15 Betaperad och vid inkomsten 12/3_2009 till KS Akuten opåverkat At, tydlig urtikaria hela kroppen, bra saturation, obetydlig rodnad men ingen påtaglig svullnad i munhålan. Got 15 Betapered when arrived to emergency department	Drug	Disorder	Indication
9	Pulm fys u a.	Finding	Body_Part	Indication

Fig. 6.1 Example of annotation using BRAT on a Swedish (anonymised) clinical text containing an adverse drug event. The example has been manually translated to English



Stockholm
University

Quizzes

- What is named entity recognition?
- What can one used it for?
- What is the difference between rule and machine learning based NER?



Stockholm
University



Questions

