

Language resources and infrastructure at TurkuNLP



TURKUNLP
.ORG



**UNIVERSITY
OF TURKU**



/ Who are we?

- TurkuNLP research group <https://turkunlp.org/>
- Main research projects/interests from past to present
 - Information Extraction / Text Mining in the bio domain
 - Turku Dependency Treebank
 - Universal Dependencies
 - Turku parser, now known as Turku Neural Parsing Pipeline
 - Large internet corpora (first Finnish, then 40+ languages)
 - Large and very large language models
 - NER
 - Paraphrase
 - Text registers + other digital language studies
 -

Topic 1: The parser



UNIVERSITY
OF TURKU

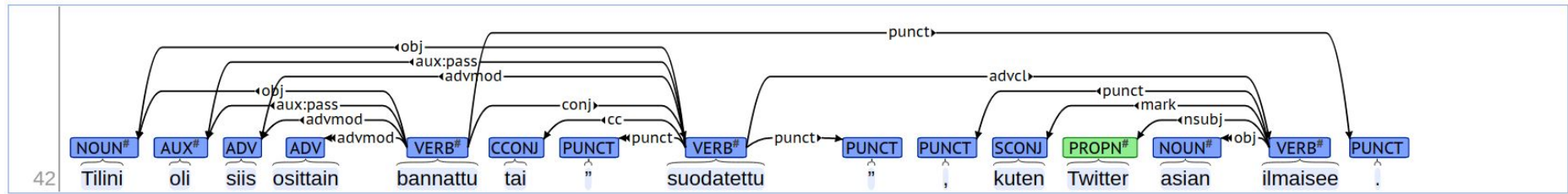


/ Treebank and parser

- Motivation (in 2009):
 - Lack of freely available tools and resources for Finnish
 - Practically nothing you could download on the Internet for Finnish
 - Training data for machine learned parser
 - Most research done on English

- Turku Dependency Treebank (2009-2014)
 - 15k sentences, 200k tokens
 - Full morphological and syntactic annotation
 - Currently lives as UD_Finnish-TDT in the [Universal Dependencies](https://universaldependencies.org/treebanks/fi_tdt/index.html) project
 - https://universaldependencies.org/treebanks/fi_tdt/index.html
 - Treebank data selection: 10 different sources (blogs, wiki, student magazines, EU legislation, ...) – A “balanced” sample
 - CC BY-SA license

/ Treebank and parser





/ Treebank and parser

```
# sent_id = b104.3
# text = Varasin pupulle ja minulle sekä sille sisarentyttärelleni, joka pääsi Turkuun lokia lukemaan, liput kaupunginteatterin Laulavat sadepisarat -musikaaliin.
1  Varasin varata VERB V Mood=Ind|Number=Sing|Person=1|Tense=Past|VerbForm=Fin|Voice=Act 0 root 0:root _
2  pupulle pupu NOUN N Case=All|Number=Sing 1 obl 1:obl _
3  ja ja CONJ C 4 cc 4:cc _
4  minulle minä PRON Pron Case=All|Number=Sing|Person=1|PronType=Prs 2 conj 1:obl|2:conj _
5  sekä sekä CONJ C 7 cc 7:cc _
6  sille se PRON Pron Case=All|Number=Sing|PronType=Dem 7 det 7:det _
7  sisarentyttärelleni sisaren#tytär NOUN N Case=All|Number=Sing|Number[psor]=Sing|Person[psor]=1 2 conj 1:obl|2:conj SpaceAfter=No
8  , , PUNCT Punct 10 punct 10:punct
9  , , PUNCT Punct 10 punct 10:punct
10 joka joka PRON Pron Case=Nom|Number=Sing|PronType=Rel 10 nsubj 10:nsubj|13:nsubj
11 pääsi päästä VERB V Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act 7 acl:relcl 7:acl:relcl _
12 Turkuun Turku PROP N Case=Ill|Number=Sing 10 obl 10:obl _
13 lokia laki NOUN N Case=Par|Number=Sing 13 obj 13:obj _
14 lukemaan lukea VERB V Case=Ill|InfForm=3|Number=Sing|VerbForm=Inf|Voice=Act 10 xcomp 10:xcomp SpaceAfter=No
15 liput lippu NOUN N Case=Nom|Number=Plur 1 obj 1:obj
16 kaupunginteatterin kaupungin#teatteri NOUN N Case=Gen|Number=Sing 19 nmod:poss 19:nmod:poss
17 Laulavat laulaa VERB V Case=Nom|Degree=Pos|Number=Plur|PartForm=Pres|VerbForm=Part|Voice=Act 18 acl 18:acl _
18 sadepisarat sade#pisara NOUN N Case=Nom|Number=Plur 19 compound:nn 17:flat:name|19:compound:nn
19 -musikaaliin musikaali NOUN N Case=Ill|Number=Sing 15 nmod 15:nmod SpaceAfter=No
20 . . PUNCT Punct 1 punct 1:punct _
```



/ The Dependency Parser

Approximate results for the analysis sub-tasks:

- Token segmentation ~99%
- POS ~98%
- Full features ~96%
- Lemmas ~95%
- Syntax ~91%



/ The Dependency Parser

- Often spotted as a text pre-processing component in Finnish industry
- Spotted in: hospitals, army, several universities, content management companies,...
- Focus has always been on highest possible accuracy, less so on SW beauty and lightweight
- Turku Neural Parser Pipeline
 - Not only Finnish, trained on Universal Dependencies
 - Combined 1st rank in the 2018 CoNLL Shared Task: 80+ treebanks
 - 1st rank on the 2020 IWPT Shared Task: 20+ treebanks into enhanced dependencies (trees + extra dependencies)



/ Go try it

https://github.com/TurkuNLP/Turku-neural-parser-pipeline/blob/master/docs/tnpp_diaparse.ipynb

https://colab.research.google.com/github/TurkuNLP/Turku-neural-parser-pipeline/blob/master/docs/tnpp_diaparse.ipynb

http://bionlp-www.utu.fi/parser_demo/

Topic 2: Neural language models



UNIVERSITY
OF TURKU



/ Neural language models

Language models based on deep neural networks pre-trained on massive text corpora are key to state-of-the-art NLP

Best-known models such as BERT and GPT introduced primarily for English by US tech giants

Massively multilingual (100+ languages) models exist, but are not competitive with dedicated monolingual ones

→ To advance Finnish NLP, we need to make our own



/ FinBERT

In 2019, we trained BERT from scratch for Finnish to make FinBERT

3B words, 110M parameters, 1M steps

FinBERT outperforms Google's mBERT and previous SOTA

Details & download (it's open):

<http://turkunlp.org/FinBERT>





/ Creating FinBERT

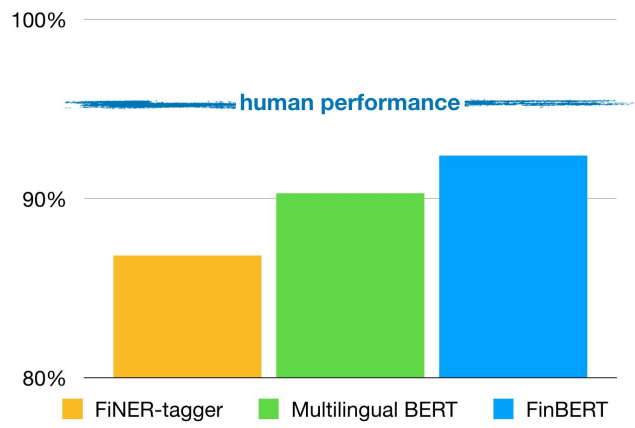
Source	Raw	Filtered
News	900M	500M
Discussion	4500M	1700M
Web crawl	8100M	1100M

Collected, cleaned and filtered all Finnish text we could find:
Yle & STT news, Suomi24 archives, Finnish internet crawl

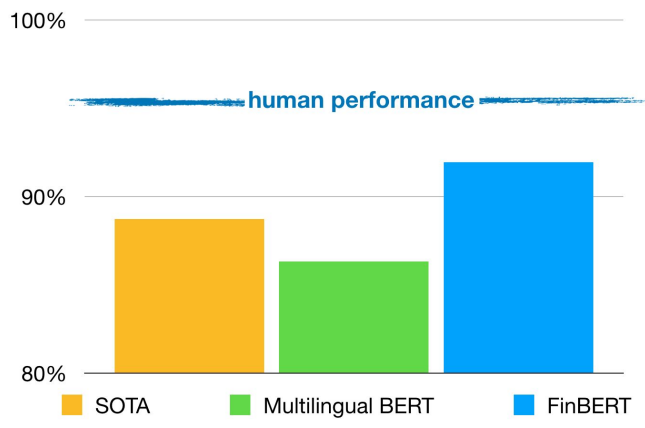
Final training data size 3.3B words, comparable to original
BERT training size

Training on Puhti at CSC

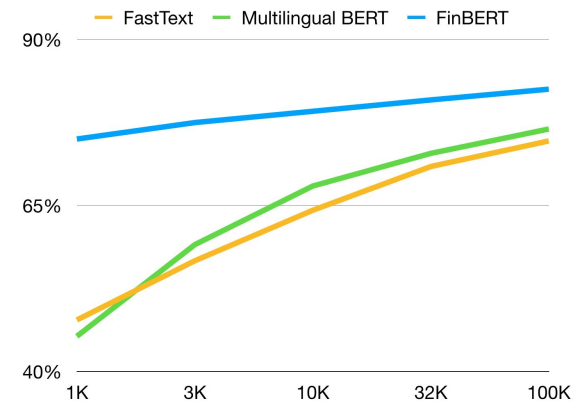
FinBERT Evaluation



Named entity recognition



Dependency parsing



Text classification

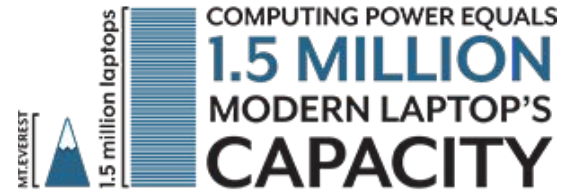
/ Generative language models for Finnish

Very large language models (10B+ parameters) for Finnish

Pilot project on LUMI, one of the world's fastest supercomputers (lumi-supercomputer.eu)

Collaboration with Jörg Tiedemann (U Hki) and the National Library of Finland

Potential applications in dialogue systems, speech recognition, question answering, and more

The LUMI logo consists of the letters "LUMI" in a large, bold, black, sans-serif font, enclosed within a double-line rectangular border.This graphic features a stack of five blue server rack icons on the left. To the right, the text reads "1 SYSTEM" in large black font, "550+ Pflop/s" in blue font, and "PEAK PERFORMANCE" in black font below it.This graphic includes a small icon of Mount Everest on the left, labeled "MTEVEREST". To its right is a tall blue bar representing "1.5 million laptops". The text "COMPUTING POWER EQUALS 1.5 MILLION MODERN LAPTOP'S CAPACITY" is displayed in blue and black fonts.The graphic shows two blue server rack icons on either side of the text "117 PB STORAGE", where "117 PB" is in a large blue font and "STORAGE" is in a smaller black font below it.



/ Generative language models for Finnish

Prompt	Voiko tietokone oppia ymmärtämään suomen kieltä?
Generated	Kuinka tietokone oppii suomen kielen? "Olemme käyttäneet paljon aikaa suomen oppisen opettamiseen ihmisille ja siinä ei ole ollut ongelmia. Kone ei osaa vielä täydellisesti suomen kielioppisääntöjä, mutta se oppii varmasti, kun aikaa on. Kuinka kauan tietokoneilla on ollut käytössä kielen oppimiseen? Aika kauan. Aikaisemmin kielioppiin liittyviä ongelmia ei ollut ja tietokone ei olisi oppinut suomenkieltä, koska se ei ymmärtäisi kieltä. Kun se nyt oppii kielen, se osaa tehdä siitä myös suomen. Voiko kielen oppiminen olla vain kielen omaksumista? Se on vain kielioppitermi, jolla on täsmällinen ja hyvä merkitys. Esimerkiksi, jos kone ei ymmärrä jotakin kieltä, vaikka se olisi kirjoitettu tietokoneella, ei tietokone ymmärrä sitä. Kielioppitermi ei vain toimi. Se ei toimi, eikä sitä ole missään. Kaikki, mitä on kirjoitettu kielen opiskelulla, on pelkkää kielen opettelua. Jos kielenoppiminen on kielen opettamista, mikä on tietokoneesi kieli? Mitä kieltä se on?

(cherry-picked from outputs generated by 0.7B parameter model trained for 72h on FinBERT training data)

Topic 3: Finnish NER



UNIVERSITY
OF TURKU



/ Finnish NER

Named entity recognition is important for numerous applications

First openly available Finnish NER corpus, FiNER, published as late as 2019 (Ruokolainen et al. 2019)

- Finnish technology news texts + additional Wikipedia test set
- Need more varied Finnish texts with NER annotation



/ Turku NER corpus

	Section	Documents	Sentences	Tokens
The TDT Finnish treebank covers a wide variety of text types	Wikipedia articles	200	2 269	31 906
	Wikinews articles	100	1 120	14 281
	University online news	50	942	13 232
	Blog entries	77	1 781	22 287
	Student magazine articles	23	1 058	14 390
Manually annotated 10,000+ mentions in 200,000+ words following FiNER guidelines	Grammar examples	80	2 002	16 982
	Europarl speeches	80	1 082	19 932
	JRC-Acquis legislation	29	1 141	23 920
	Financial news	50	1 002	12 477
	Fiction	65	2 739	32 709
	Total	754	15 136	202 116

<https://github.com/TurkuNLP/turku-ner-corpus>



/Turku NER corpus

PERSON Erik Justander (**DATE** noin 1623 , **LOCATION** Turku – **DATE** 10. marraskuuta 1678 , **LOCATION** Mynämäki) oli kirkkoherra ja **ORGANIZATION** Turun akatemian runousopin professori .

PERSON Justander varttui luultavasti kasvattipoikana raatimies **PERSON** Henrik Tavastin perheessä ja pääsi ylioppilaaksi **DATE** vuoteen 1645 mennessä .

Valmistuttuaan filosofian maisteriksi **DATE** 1653 hän toimi kreivi **PERSON** Johan Oxenstiernan kirjastonhoitajana **DATE** 1654 ja **ORGANIZATION** Turun akatemian runousopin

professorina **DATE** vuosina 1655-1667 .

Six annotated types: person, organization, location, date, product, event



/ Turku OntoNotes Entities Corpus

PERSON (**DATE** **GPE**) on **NORP** suomalainen jalkapalloilija , joka pelasi kaudella **DATE** **ORG** 2007 PK-35:ssä .

WORK OF ART A Day at the Races on vuonna **DATE** 1976 julkaistu **NORP** brittiläisen rock-yhtye **ORG** Queenin **ORDINAL** viides albumi .

FACILITY Kajaanin linna on **DATE** 1600-luvulla rakennettu linnoitus **LOCATION** Ämmäkosken saarella **LOCATION** Kajaaninjoella nykyisen Kajaanin kaupungin keskustassa . **GPE**

FiNER and Turku NER corpus annotations compatible, but restricted and non-standard → combined, revised annotation from 6 to 18 OntoNotes types

Cross-lingually compatible with English OntoNotes NE corpus

<https://github.com/TurkuNLP/turku-one>



/ Evaluation results

Best results on all corpora for
FinBERT

Turku NER corpus + FiNER:
93.7% F-score (6 types)

Turku OntoNotes Entities corpus:
93.0% F-score (18 types)

cf. OntoNotes English NE SOTA:
92.1% F-score

Type	Prec.	Rec.	F-score
PERSON	94.12	97.15	95.60
NORP	94.63	96.15	95.36
FAC	67.83	40.00	50.23
ORG	94.14	94.06	94.10
GPE	95.33	97.36	96.33
LOC	87.12	86.50	86.78
PRODUCT	87.53	88.08	87.81
EVENT	72.17	79.46	75.59
WORK_OF_ART	75.00	77.33	75.97
LAW	90.83	96.74	93.69
LANGUAGE	93.05	95.00	94.01
DATE	94.70	94.78	94.74
TIME	81.70	84.32	82.98
PERCENT	95.60	98.61	97.08
MONEY	95.36	94.79	95.08
QUANTITY	87.18	90.90	89.00
ORDINAL	90.33	91.37	90.84
CARDINAL	94.01	95.36	94.68



/ Finnish NER: demo and code

- Online demo & code: <https://turkunlp.org/fin-ner.html>
- Direct link to demo: <http://86.50.253.19:8001/tagdemo/>
- We can run it also in Google Colab:

https://github.com/fginter/ainl_2020_tutorial/blob/main/turku_ner.ipynb

- **Spotted very often in active use in Finnish data science industry**

Topic 4: Online registers



UNIVERSITY
OF TURKU



/ Text registers (~genres)

- Crawled corpora are an important part of NLP and linguistic research
- Dumping all crawled data onto a single pile is not sufficient
- Register is the most important factor explaining linguistic variation
- Attaching register information to crawled corpora would allow for more structured usage of the data

34 PRIDE AND PREJUDICE.
"I am no longer surprised at your knowing *only* six accomplished women. I rather wonder now at your knowing *any*."
"Are you so severe upon your own sex as to doubt the possibility of all this?"
"I never saw such a woman with so little capacity, and taste, and application as you describe, united."
Mrs. Hurst and Miss Bingley were not less sensible of the injustice of her implied censure, than of the testing that they knew many of their own sex were capable of. Mrs. Hurst's bitter complaints of their inferiority were now put forward. As all conversation turned upon Elizabeth, Elizabeth soon afterwards left the room.

Ingredients

- 3/4 cup granulated sugar
- 3/4 cup packed brown sugar
- 1 cup butter, softened
- 1 teaspoon vanilla
- 1 egg
- 2 1/4 cups Gold Medal™ all-purpose flour
- 1 teaspoon baking soda
- 1/2 teaspoon salt
- 1 package (12 ounces) semisweet chocolate chips (2 cups)
- 1 cup coarsely chopped nuts, if desired

Article Talk

Borgio Verezzi

From Wikipedia, the free encyclopedia

Borgio Verezzi (Ligurian: *Bôrzi Veresso*) is a *comune* (municipality) in the Province of Savona in the Italian region Liguria, located about 60 kilometres (37 mi) southwest of Genoa and about 20 kilometres (12 mi) southwest of Savona.

Contents [hide]

- 1 Geography
- 2 Main sights
- 3 References
- 4 External links



/ Online register corpora

- Samples from the *unrestricted* web
 - Full range of registers on the web, not just selected classes!
 - Manually annotated for *register*
 - Allow to identify registers in crawled datasets + linguistic register analysis
- Challenges
 - Which registers to include? What kinds of texts do we have in the Internet?
 - Register classes not always well defined, after all, the Internet is a jungle with no rules on how to write!
- Our register scheme follows the English CORE taxonomy developed by Biber & Egbert



/ Register classification

Narrative: News reports/News blogs, Sports reports, Personal blog, Historical article, Short story / Fiction, Travel blog, Community blog, Online article

Informational Description: Description of a thing, Encyclopedia articles, Research articles, Description of a person, Information blogs, FAQs, Course materials, Legal terms / conditions, Report

Opinion: Reviews, Personal opinion blogs, Religious blogs/sermons, Advice

Interactive discussion: Discussion forums, Question-Answer forums

How-to/instructional: How-to/instructions, Recipes

Informational persuasion: Description with intent to sell, News+Opinion blogs/Editorials

Lyrical: Songs, Poems

Spoken: Interviews, Formal speeches, TV transcripts



/ Our CORE datasets

FinCORE 10,754 documents

SweCORE 3,900 documents

FreCORE 3,900 documents

- + Small *evaluation* collections in many more: Chinese, Arabic, Japanese, German, Spanish, Turkish

All available at github.com/TurkuNLP



/ Automatic register labeling

- Best multilingual models 71% F1-score
- For Finnish 78% F1-score
- For English 69 % F1-score

huggingface.co/datasets/mhtoin/register_oscar/tree/main



Hugging Face

Search models, datasets, users...



Dataset: mhtoin/register_oscar



like

0



TURKUNLP
.ORG



mhtoin

Upload sw.jsonl with git-lfs

6959d00



.gitattributes

1.19 kB



ar.jsonl

173 GB



en.jsonl

1,233 GB



es.jsonl

162 GB



fr.jsonl

86 GB



hi.jsonl

17 GB



pt.jsonl

71.8 GB



sw.jsonl

9.29 MB



ur.jsonl

4.65 GB



zh.jsonl

342 GB



Topic 5: Other resources



UNIVERSITY
OF TURKU



/Turku paraphrase corpus

- Brand new corpus
- 100,000+ paraphrase pairs
- All manually selected and classified
 - Unique in giving also document context for most of the pairs
 - The largest corpus of this kind for any language
- Data:
 - Movie subtitles
 - Aligned news from different sources
 - Student exam answers to the same question
 - Different translations of the same texts by students

<https://turkunlp.org/paraphrase.html>



/ Perfect example

Paraphrase pairs with minimal lexical overlap

‘Pohdi haastattelujen positiivisia ja negatiivisia puolia.’:

...Haastatteluja voi olla vaikea vertailla keskenään tai laittaa järjestykseen, jos ne ovat kovin erilaisia toisistaan (esimerkiksi strukturoimattomissa haastatteluissa). **Ryhmähaastattelussa vaarana on, että osa puhuu liikaa ja osa ei saa ollenkaan suunvuoroa.** Haastateltavalla pitäisi aina olla luottavainen ja turvallinen olo, jotta haastattelusta saadaan kaikki irti...

...Jos on kyseessä ryhmähaastattelu, niin osallistujat saattavat puhua toistensa päälle, jolloin on hankala saada selvää. Henkilön elekielen oikea tulkinta voi olla haastavaa. **Ryhmähaastattelussa ujoimmat saattavat jäädä ilman ääntä ja vahvimmat henkilöt jyrätä omalla mielipiteellään.**

...Ääripäät jäävät taustalle ja lopputulos voi olla tylsä konsensus. **Ei päästä esiin hiljaisempia ja ujompia yksilöitä vaan äänekkäimmät nousevat esille.** Keskustelu voi karata liikaa jos moderaattori on on täysin ulkona keskustelun ohjaamisessa, vaikka toki roolin tulee olla lähtökohtaisesti suhteellisen näkymätön...



/ Paraphrase modelling

- Given a pair of statements, tell how close in meaning they are
- Deep neural network models, go beyond surface
- Can be used (among other tasks) to query for paraphrases
- Here: search space of 400,000,000 sentences

Query: **Sain potkut.**

Hit: Minut sanottiin irti.
Hit: Minä Sain potkut.
Hit: Minä annoin potkut.
Hit: Hän potki minut pihalle.
Hit: Minut heitettiin pihalle.
Hit: Hän heitti minut pihalle.
Hit: Hän sanoi minut irti.
Hit: Laitoin sen pois.
Hit: Minä otin lopputilin.
Hit: Minäkin sain kenkää.
Hit: Minut potkittiin pihalle.
Hit: Hän antoi minulle kenkää.
Hit: Hän potki minut pellolle.
Hit: Minut heitettiin pois.
Hit: Otin sen sivuun.
Hit: Panin sen pois.
Hit: Annoin sinulle potkut.
Hit: Jouduin antamaan lopputilin.
Hit: Annoin sinulle kenkää.
Hit: Minut potkittiin pois.
Hit: Minut piti heittää pihalle.

Query: **Turussa on kivaa asua, tykkään tästä paikasta.**

Hit: Turussa on kiva ollut asua, tykkään tästä kaupungista.
Hit: Minusta Turussa on aivan ihana asua!
Hit: Mutta Turku on kyllä ihana kaupunki, täällä on kiva asua ;)
Hit: Viihdyn ihan kauhean hyvin täällä Turussa.
Hit: Pidän Turusta, koska turkulaiset ovat minua kohtaan tosi mukavia.
Hit: Jeps, tuo on juurikin se Turun kämppä Tykkään itekkin!
Hit: Kiitos, Turku on vaikuttanut kyllä kivalta paikalta asua! :)
Hit: Viihdyn erinomaisesti Turussa ja pidän sitä kotikaupunkinani nykyään.
Hit: Ja ihanaa, että asut Turussa, minäkin asun.
Hit: Kaupunkina tykkään Turusta tosi paljon.
Hit: Turussa on kyllä niin mahtava asua.
Hit: – Olen viihtynyt Turussa hyvin.
Hit: Turussa asumisessa on hyvät puolensa.
Hit: Turussa on kivojakin asuinalueita ;)
Hit: Onni on asua Turussa.
Hit: Onni on asua Turussa...
Hit: Turku on ollut hyvä paikka asua ja täällä on mukavia ihmisiä.
Hit: Olen viihtynyt Turussa hyvin.
Hit: Tykkäsin tosi paljon, ja asuu nykyään siis Turun lähetyillä.
Hit: Nyt minulla on kuitenkin Turussa kiva asunto.
Hit: Just nyt on aika ihanaa asua Turussa.



/ Turku sentiment corpus

- Broad-coverage corpus annotated for sentiment at multiple layers
- e.g. author emotion vs. assessment of things mentioned in text
 - “I’m delighted that the new iPhone is so bad”
- Fully manual annotation with multiple trained annotators
- Covering 10 genres of TDT corpus plus additional sources such as Finnish twitter



/ Wikidata normalization corpus

Turun yliopisto (lyhenne TY tai UTU) on ensimmäinen täysin suomenkielinen yliopisto, joka perustettiin 1920. Elokuussa 2019 yliopiston rehtorina aloitti Jukka Kola. Yliopiston viimeisenä kanslerina

University of Turku (Q501841)

university in Turku, Finland

Turun yliopisto | Åbo universitet | Turun University

- Linking of named entities in the Turku NER corpus to Wikidata
- Support for linking mentions in text to structured resources
- Tasks: candidate generation + context-based disambiguation

Conclusions



UNIVERSITY
OF TURKU



/ Conclusions

- “Classical” tasks around morphosyntactic parsing (syntax, tagging, lemmatization) on clean text are very close to human performance, from an engineering point of view “solved”
- Deep learning has changed NLP; FinBERT is now a component of state-of-the-art methods for many Finnish tasks
- New and upcoming resources for numerous harder tasks within the realm of natural language understanding and content classification, these will be the next research frontier for Finnish NLP
- All datasets and resources for Finnish fully open and usable by anyone
- Don’t expect that others will make your data for you, if you deal with a small language like Finnish