Creating annotated corpora

Jelke Bloem & Giovanni Colavizza

Text Mining
Amsterdam University College

April 21, 2023

Announcements

- Tuesday 18/04: Assignment 3
- Tuesday 18/04: Reading assignment - BERT
- Thursday 20/04: Project update 0 (repository+proposal)

Overview

1. NLP and Human Annotations

2. Linguistic Annotation
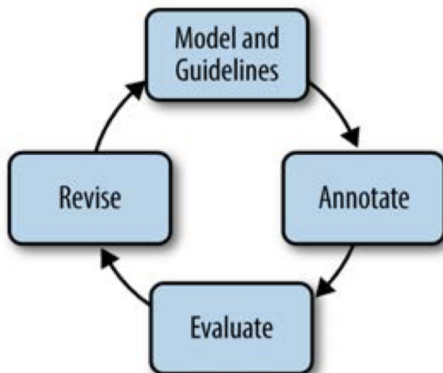
3. Evaluation

4. Text corpora

5. Platforms and shared tasks
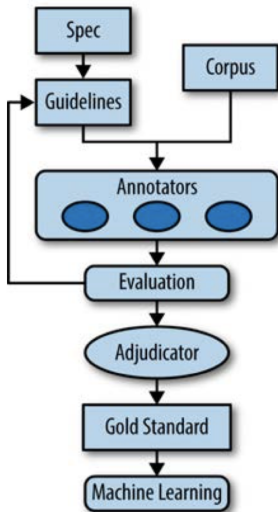
NLP and Human Annotations

NLP and Human Annotations

- NLP (and ML in general) is driven by human-annotated corpora.
- http://nlpprogress.com.
- Annotation is **difficult** and **expensive**.

Annotation pipeline

Pustejovsky and Stubbs, 2012. Natural Language Annotation for Machine Learning.

Annotation pipeline

Pustejovsky and Stubbs, 2012. Natural Language Annotation for Machine Learning.

Specs and guidelines

- Goal: given our problem, *how can we formalize our description of the annotation process for multiple annotators to provide the same judgment?*
  - What is the goal of the project?
  - How will the annotation be created? (For example, which tags or documents to annotate first, how to use the annotation tools, etc.)
  - What is each tag called and how is it used? (provide examples and discuss problematic choices.)
  - What parts of the text do you want annotated?
- Note: annotation is usually boring and time-consuming, and cannot be done for 8 hours straight. Annotators also get better over time: early annotations might be discarded.

Pustejovsky and Stubbs, 2012. Natural Language Annotation for Machine Learning.

Adjudication

- **Adjudication** is the process of deciding on a single annotation for a piece of text, using information from all independent annotators.
- Yes, it is only possible when multiple annotators independently annotate (at least some) of the corpus. *This is a very good procedure to follow, and the only one which will allow to evaluate results*.
- It can be as time-consuming (or more so) as a primary annotation.
- It does not need to be identical with a primary annotation (all annotators can be wrong by chance), but unlikely so.

Pustejovsky and Stubbs, 2012. Natural Language Annotation for Machine Learning.

Automatic annotation

- **Manual annotation**: Data is annotated by
  - Experts
  - The crowd (e.g. Amazon Mechanical Turk)
- Based on:
  - A (expert-created) ground truth
  - Annotation guidelines
  - Elicitation of implicit/explicit knowledge
- **Semi-automatic annotation**: A computer program is used to annotate the data, and annotators perform checks and corrections
- **Automatic annotation**: A computer program (predictive model, parser etc.) is used to annotate the data

Automatic annotation

Advantages

- Process far more data
- Study rare phenomena
- Estimate probabilities more accurately
- Flexible: work with your own type of text
- Don't need to hire a bunch of students to annotate...!

Disadvantages

- Annotation errors
- Biases of automatic system are introduced
- Cannot really be used as training data for machine learning
- Annotation may constrain what can be researched

Automatic annotation: Potential biases

- Random errors and systematic errors
- More errors for rare phenomena
- More errors when there is more ambiguity
- More errors for larger structures, longer sentences
- Multi-word units / idiomatic expressions
- More errors for out-of-domain data
- More errors when original text contains errors

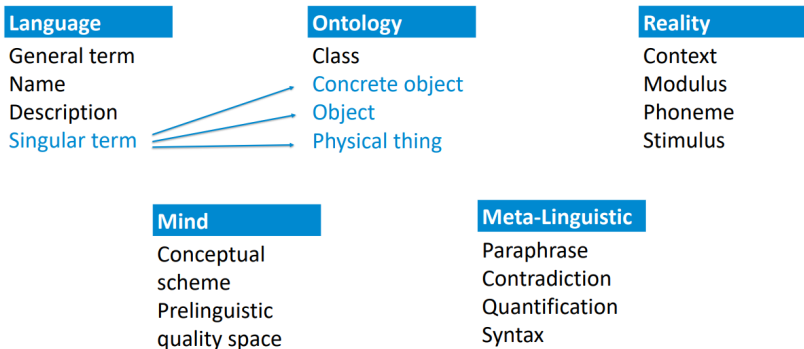# Automatic annotation: Checking quality

- Find existing evaluations on similar text types
- Manually check (parts of) text

When querying based on automatic annotation:

- Manually check results (precision)
- Check on the basis of a simpler layer of annotation
- Check on the basis of exemplars (recall)

Ground truth

- Quine's view of the world

Language
General term
Name
Description
Singular term

Ontology
Class
Concrete object
Object
Physical thing

Reality
Context
Modulus
Phoneme
Stimulus

Mind
Conceptual
scheme
Prelinguistic
quality space

Meta-Linguistic
Paraphrase
Contradiction
Quantification
Syntax

Eliciting explicit knowledge

- Distributional semantic modeling of Quine

Synonym detection task

What word is most related to 'Information' ?

a) Learning	b) Reductions
c) Collateral	d) Application
e) Ordered Pair	f) None of these words is even remotely related

Coherence task

What word does not belong to the group?

a) Numbers	b) Pronouns
c) Subtraction	d) Actually
e) No coherent group can be formed from these words	

FIGURE 1: *Target word a), nearest neighbours b) and c), and outlier d).*

Linguistic Annotation

Levels of linguistic annotation

- Tokenization, lemmatization…
- Part-of-speech tagging

```
>>> text = word_tokenize("They refuse to permit us to obtain the refuse permit")
>>> nltk.pos_tag(text)
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit', 'VB'), ('us', 'PRP'),
('to', 'TO'), ('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN'), ('permit', 'NN')]
```

|			
CC	Coordinating conj.	TO	infinitival *to*
CD	Cardinal number	UH	Interjection
DT	Determiner	VB	Verb, base form
EX	Existential there	VBD	Verb, past tense
FW	Foreign word	VBG	Verb, gerund/present pple
IN	Preposition	VBN	Verb, past participle
JJ	Adjective	VBP	Verb, non-3rd ps. sg. present
JJR	Adjective, comparative	VBZ	Verb, 3rd ps. sg. present
JJS	Adjective, superlative	WDT	Wh-determiner
LS	List item marker	WP	Wh-pronoun
MD	Modal	WP\$	Possessive *wh*-pronoun
NN	Noun, singular or mass	WRB	Wh-adverb
NNS	Noun, plural	#	Pound sign
NNP	Proper noun, singular	\$	Dollar sign
NNPS	Proper noun, plural	.	Sentence-final punctuation
PDT	Predeterminer	,	Comma
POS	Possessive ending	:	Colon, semi-colon
PRP	Personal pronoun	(	Left bracket character
PP\$	Possessive pronoun	)	Right bracket character
RB	Adverb	"	Straight double quote
RBR	Adverb, comparative	'	Left open single quote
RBS	Adverb, superlative	"	Left open double quote
RP	Particle	'	Right close single quote
SYM	Symbol	"	Right close double quote

Levels of linguistic annotation

- Syntax: Dependency parsing
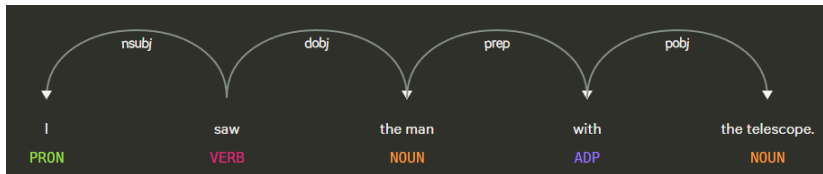
Levels of linguistic annotation

- Syntax: Constituency parsing
- Syntactically annotated corpora are also called treebanks
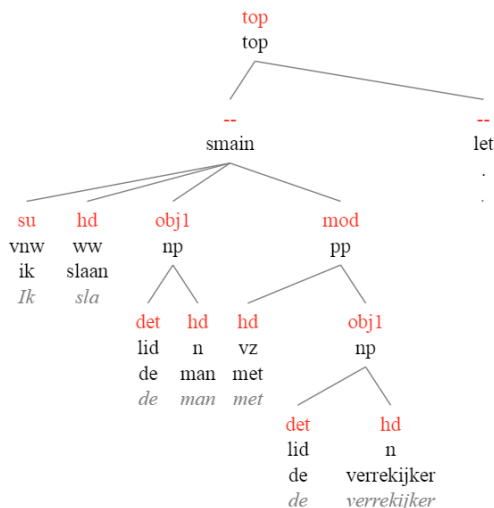
Levels of linguistic annotation

- Semantic: DBPedia (linking to ontology/knowledge base)

Levels of linguistic annotation

- Discourse: Rhetorical Structure Theory
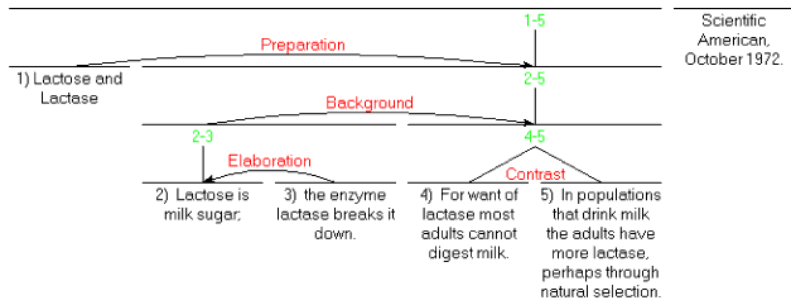
Evaluation

Interannotator agreement

annotator A

	puppy	fried chicken
puppy	6	3
fried chicken	2	5

annotator B

observed agreement = 11/16 = 68.75%

Credit: David Bamman, UCBerkeley.

Cohen's kappa

- Similar idea to mutual information: observed minus expected agreement.
- Cohen's kappa is defined for two annotators over the same set of annotation tasks:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Where p_o is the observed correct agreement and p_e the expected correct agreement.

annotator A

	puppy	fried chicken
puppy	7	4
fried chicken	8	81

annotator B

Credit: David Bamman, UCBerkeley.

Cohen's kappa example

- $p_o = 0.88$
- $p_e = P(A = puppy)P(B = puppy) + P(A = chicken)P(B = chicken)$
-
$$\kappa = \frac{0.88 - 0.773}{1 - 0.773} = 0.471$$

annotator A

Credit: David Bamman, UCBerkeley.

Cohen's kappa scores

Note: these are rules of thumb.

0.80-1.00	Very good agreement
0.60-0.80	Good agreement
0.40-0.60	Moderate agreement
0.20-0.40	Fair agreement
< 0.20	Poor agreement

Exercise: try to calculate fringe cases. E.g., 50/50 puppy/chicken all in agreement, 0/100 puppy/chicken all in agreement, 50/50 wrong puppy/chicken all in agreement.
Credit: David Bamman, UCBerkeley.

Fleiss' kappa

- Extension to multiple annotators (> 2).
- Defined as Cohen's kappa but comparing pairs of annotators:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

annotator A

	puppy	fried chicken
puppy	7	4
fried chicken	8	81

annotator B

Credit: David Bamman, UCBerkeley.

Fleiss' kappa

- Number of annotators who assign category j to item i: n_{ij}.
- For item i with n annotations, how many annotators agree among all $n(n-1)$ possible pairs:

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} n_{ij}(n_{ij} - 1)$$

- Note that N is the number of items, and K the available annotation categories. Average agreement among all items:

$$P_o = \frac{1}{N} \sum_{i=1}^{N} P_i$$

Fleiss' kappa, continued

- Probability of category j:

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}$$

- Expected agreement by chance:

$$P_e = \sum_{j=1}^{K} p_j^2$$

- Back to original formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Text corpora

Corpus linguistics

- Balanced corpus
- Learner corpus
- Historical corpus
- Parallel corpus
- Spoken corpus (transcribed)
- N-gram corpus

Metadata

- Information about the texts in a corpus
- Year of publication, author, medium, register, edition, chapter, age of speaker, language, encoding, size etc.
- Particularly important for historical corpora and corpora where distinct documents matter (academic texts, movie reviews)

Corpus creation: Things to note

- Have clear selection criteria
  - Avoid subjective choices/criteria ('many errors')
- Select representative texts for the topic
- Take a balanced sample (if needed for e.g. training purposes)
- Think about copyright issues
- Consider availability and format
  - On paper
  - Images/scans
  - Proprietary format
  - Plain text
  - Annotated text

Corpus analysis

- Keyword-in-Context

```
              an accident waiting to happen – Idioms by The Free Diction
                 accident waiting to happen Definition from Wiktionary,
. News 'This is a Ferguson waiting to happen:' Activists speak out agains
e > Idioms > A > Accident waiting to happen Idiom: Accident waiting to
              Trouble Waiting to Happen From Wikipedia, the free enc
 26 April 2015. 'Nightmare Waiting to Happen': Quake Experts Gathered in
rt Start reading Accidents Waiting to Happen on the free Kindle Reading A
 Earthquake Was "nightmare Waiting to Happen" Slate Sign In Sign Up Slate
gions Disasters 'Nightmare waiting to happen': Experts gathered in Nepal
an></span></div> A tragedy waiting to happen By Colin Stark Updated 1759
bal quake was a "nightmare waiting to happen" Breaking News US Secretary
Nepal quake was 'nightmare waiting to happen' Major earthquakes hit Nepal
mph of Life Contact Tarot Waiting to Happen – A Tarot Deck by Andrew McG
ergy Solution or Accident Waiting to Happen: The Public and Nuclear Powe
y Angela Weight and Sanity Waiting to Happen don't live here anymore. Cor
 of Things Is a Revolution Waiting to Happen The challenge of the IoT is
arthquake was a "nightmare waiting to happen" says lead scientist Experts
       Angela Weight Sanity Waiting to Happen Skip to content Home Book(s)
eos Games Music En Deadman Waiting To Happen by nProcess Random Animation
             an accident waiting to happen – definition in the British
|Bhuj-temblorA catastrophe waiting to happen? The author has posted comme
grade Winter Meeting trade waiting to happen between M's, Rockies JP Morc
e Author Says Disaster Was Waiting to Happen Don't Miss Out — Follow us o
 earthquake was a disaster waiting to happen Sitting on one of the most e
gallery 10 Sports Injuries Waiting to Happen of Advertisement Skip this a
en Noun (plural disasters waiting to happen) Something potentially very
st Classic FAILs accident waiting to happen Share on Facebook- Featured
```

Corpus analysis

- Keyword-in-Context
- Collocations

+ NOUN		NEW WORD	?
1334	3.04	job	
945	2.37	point	
870	3.73	article	
820	2.32	book	
769	4.02	example	
741	2.11	work	
632	3.33	post	
596	3.31	choice	
508	3.01	star	
455	2.99	opportunity	
406	2.76	source	
374	3.04	quality	
363	2.92	performance	
331	2.61	condition	
300	2.20	chance	
299	2.80	skill	
264	2.08	piece	
250	2.51	resource	
214	2.67	tool	
212	3.80	rating	
201	2.50	review	

Corpus analysis

- Keyword-in-Context
- Collocations
- Collostructions
  - ? waiting to happen

Prediction	Prob.	Prob. Top-K
event	0.097763	0.2703
disaster	0.064560	0.1611
accident	0.059664	0.1394
explosion	0.049361	0.1390
invasion	0.016694	0.0486
earthquake	0.016525	0.0478
action	0.016206	0.0422
emergency	0.014662	0.0417
attack	0.013799	0.0403
miracle	0.013404	0.0371
adventure	0.011491	0.0324

Querying corpora

- Nederlab with Corpus Query Language

Querying corpora

- Nederlab with Corpus Query Language

Historie van mejuffrouw Sara Burghart
datering: 1782
auteur: **Aagje Deken (Amstelveen (Noord-Holland), 1741-Den Haag (Zuid-Holland), 1804) Betje Wolff (Vlissingen (Zeeland), 1738-Den Haag (Zuid-Holland), 1804)**
genre: **fictie, proza, briefroman**
collectie: **DBNL**
aantal hits: 1

| dat | zy | evenwel | beter | de | klink | kan | maken | dan | Jaantje | en |
| dat | zy | evenwel | goed | de | klinken | kunnen | maken | dan | jaan | en |

Kransje van letter-bloempjes, voor Neerlandsch jufferschap
datering: 1790
auteur: **Gerrit Manheer (Rotterdam (Zuid-Holland), 1749-Rotterdam (Zuid-Holland), 1807)**
genre: **fictie, poezie, liederen/liedjes**
collectie: **DBNL**
aantal hits: 1

| Die | Gij | , | door | lachjes | klink | in | uw | ' | fluweele | boeijen |
| die | gij | . | door | lach | klink | in | uw | ' | fluweele | boeijen |

Werken van het Amsteldamsch Dicht- en Letteröfenend Genootschap. Te Amsteldam by M. de Bruyn, 1790. In gr. 8 vo. , 118 bladz .
uit: Vaderlandsche letteroefeningen. Jaargang 1791
datering: 1791
genre: **non-fictie, periodieken, tijdschriften, jaarboeken, letterkunde (secundair)**
collectie: **DBNL**
aantal hits: 1

| Dat | den | der | Barden | klaagzang | klink | ' | , | Hoe | Elza's | jammer |
| dat | de | de | bard | klaagzang | klink | ' | . | hoe | Elza | jammer |

Querying corpora

- NLCOW14 (web corpus)

... NLCOW14AX01	[pos="verbpapa"%c][word="kunnen"%c][word="hebben"%c] 0	100 Preview again Export results Delete	

Wel , it ' s a hell of a job , zou wijlen Fortuyn	gezegd kunnen hebben	.	http://www.repu...
Zij zou ook op mij	geschoten kunnen hebben	in dat restaurant .	http://home.hcc...
Uit de stukken blijkt verder dat [A] , [B] en [eiser] zelf kort na het ongeval naar het ziekenhuis zijn vervoerd en daarom niet met medeleerlingen van [B]	gecommuniceerd kunnen hebben	.	http://www.lets...
" en hij zegt dan verder , dat die andere mensen het niet	gedaan kunnen hebben	omdat zij niet gearresteerd waren .	http://www.hebr...
" Ik haal een Grieks woord om de context te verduidelijken . de laatste zin die begint met " met andere woorden " zou zich zo in Korinthe	afgespeeld kunnen hebben	.	http://forum.re...
Anders zou zich uit de intensiteit van du Perrons gevoelens een tragedie	ontwikkeld kunnen hebben	, die de Hamlet zou hebben aangevuld en in tragische kracht wellicht overtroffen .	http://www.dbnl...
Dostojevski zou het	geschreven kunnen hebben	.	http://m.nrc.nl...

Querying corpora

- Word relation search with PaQu

Querying corpora

- Syntactic search with PaQu

corpus: Lassy Groot: NL-wiki — 8 707 708 zinnen

XPATH query (voorbeelden):

```
//node[@cat="ssub" and node[@rel="hd" and @root="heb" and @pos="verb" and
number(@begin) < number(../node[@rel="vc" and @cat="ppart"]/node[@rel="hd" and
@pos="verb"]/@begin)] and node[@rel="vc" and @cat="ppart" and node[@rel="hd" and
@pos="verb"]]]
```

aantal: 20

Zoeken   Wissen   Reset

1. Hij werd echter reeds in januari 1931 lid van de NSDAP nadat hij in december 1930 een toespraak van Adolf Hitler had bijgewoond in de Berlijnse Hasenheide . ✜
2. Zelf schreef hij dat hij een maand had getwijfeld , maar dat hij uiteindelijk toch besloten had om lid te worden , omdat Hitler helemaal niet stereotiep was overgekomen in de toespraak . ✜
3. Van alle veroordeelde nazi-kopstukken was hij de enige die schuld had bekend . ✜
4. In werkelijkheid waren de verbeteringen van slaap- en verblijfsomstandigheden van dwangarbeiders al gepland voordat Speer de fabrieken had bezocht waar deze werkten . ✜
5. Later kwam tevens aan het licht dat hij had medegewerkt aan de uitbreidingsplannen voor Auschwitz . ✜
6. Die gevoelsarmoede , de afwezigheid van normale menselijke reacties , is de belangrijkste onbeantwoordbare vraag die Fest lang bezig heeft gehouden . ✜
7. In 1999 werd Agassi de vijfde speler in de geschiedenis van de sport die alle vier de Grand Slam toernooien had gewonnen : de Australian Open , de Open Franse Tenniskampioenschappen , Wimbledon en de US Open . ✜

Querying corpora

- Example-based syntactic search with GrETEL

GrETEL (Augustinus, Vandeghinste, and Van Eynde 2012)

■ Syntactic search: Querying by example

ik denk dat ik het heb begrepen

sentence	Ik	denk	dat	ik	het	heb	begrepen	.
word	○	○	○	○	○	○	○	○
lemma	○	○	○	○	○	●	○	○
word class	○	○	○	○	○	○	●	○
optional in search	◉	◉	◉	◉	◉	○	○	◉

OPTIONS

☑ Respect word order

ssub

hd   vc
verb ppart
heb

hd
verb

Platforms and shared tasks

Annotation tools

- Brat `http://brat.nlplab.org`
- Inception `https://inception-project.github.io`
- Prodigy `https://prodi.gy`

Annotation platforms

- Supervisely `https://supervise.ly`
- Dataturks `https://dataturks.com`
- Amazon Mechanical Turk `https://www.mturk.com`
- Figure Eight `https://www.figure-eight.com`
- AIcrowd `https://www.aicrowd.com`