Information extraction & automated knowledge graph construction

Lecture 5: IE annotations

Lauriane Aufrant (Inria Paris)

Université Paris Cité – UFRL – M2 LI – 07/12/2023

Progress

- ~~09/11 – Overview of information extraction~~
- ~~16/11 – Entities, relations...~~
- ~~23/11 – Coreference, linking...~~
- ~~30/11 – From IE to automated knowledge graph construction~~
- 07/12 – IE annotations
- 14/12 – IE in a specialty domain

Today: IE annotations

- Annotation guidelines
- The annotation process
- Good practices for corpus annotation
- **The project**

Annotation guidelines

The annotation scheme: both a convention and a guide

- Main guidance given to annotators
- Provided with the corpus to help understand its annotations
- Just indicating the task does not suffice: too many variants
- Defines the syntax of annotations (format, vocabulary...) and their semantics (scope of classes, encompassed cases, decision criteria...)

Example (outside IE): Universal Dependencies guidelines

UD Guidelines

- Basic principles
  - Tokenization and word segmentation
  - Morphology
  - Syntax
  - Enhanced dependencies
  - CoNLL-U format and its extensions
  - Typos and other errors in underlying text
- Annotation guidelines
  - Nominals
  - Simple clauses
  - Complex clauses
  - Comparative constructions – working group materials
  - Other constructions
- Documentation of tags, features and relations
  - POS tags (single document)
  - Features (single document)
    - Layered features
    - Features in data (list of **all** features and values used in treebanks, including those that are **not defined** by the universal guidelines)
  - Syntactic relations (single document)
    - Relations in data (list of **all** relation subtypes that are used in treebanks)
  - Conversion from other tagsets to UD tags and features
  - MISC attributes
- Incubator for Construction-Oriented Documentation (it will be moved here when it is mature enough)

This is the online documentation of UD guidelines v2 (launched 2016-12-01 with subsequent revisions). For change history, see Guidelines Changes.

Tokenization and Word Segmentation

Words are generally delimited by whitespace or punctuation. No tokens in any of the UD English corpora currently contain whitespace. Multiword tokens should be used for English clitics, such as *'ll* (reduced form of the auxiliary *will*), *_n't* (reduced form of *not*) and *'s* (possessive clitic). For example, *don't = do + n't*). As of mid 2021, multiword tokens are used in the following English corpora: GUM, GUMReddit, and EWT, are partially used in ParTUT (used for forms like *ain't* and *can't* but not for forms that are concatenative like *John's* or *she'll*), but are not used in: PUD, LinES, Pronouns, or ESL. If multiword tokens are not present, clitics in English can usually be identified by using the `SpaceAfter=No` annotation and it also allows distinguishing between otherwise identical token sequences, such as "can not" versus "cannot".

Units that should be regarded as separate syntactic words include:

- Clitic auxiliaries ('ll, 'm, 's, 've, 'd, ...)
- Possessive genitive markers ('s, ')
- Clitic negation (n't, and also not in cannot)
- Most hyphenated terms (search-engine becomes 3 words: search, -, engine)

Units that are not tokenized apart include:

- Acronyms (FBI, U.S.)
- Abbreviations without spaces (e.g., i.e.)
- Some hyphenated words, with common prefixes or occasionally suffixes, such as *e-mail* or *co-ordinated*

Example (in IE): ACE guidelines

3 Entity Types and Subtypes

3.1 Persons (PER)

Each distinct person or set of people mentioned in a document refers to an entity of type Person. For example, people may be specified by name ("John Smith"), occupation ("the butcher"), family relation ("dad"), pronoun ("he"), etc., or by some combination of these. Dead people and human remains are to be recorded as entities of type Person. So are fictional human characters appearing in movies, TV, books, plays, etc.

There are a number of words that are ambiguous as to their referent. For example, nouns, which normally refer to animals or non-humans, can be used to describe people. If it is clear to the annotator that the noun refers to a person in a given context, it should be marked as a Person entity.

He is [a real turkey] [1]

[The political cat of the year]

She's known as [the brain of the family]

3.1.1. Subtypes for Person

We will further classify Person entities with the following subtypes.

PER.Individual

If the Person entity refers to a single person, tag it as PER.Individual.

[Bill Clinton]

[Edmund Pope]

[The President of the U.S.]

The police found [[his] body]

Writing annotation guidelines

- Need to cover all possible cases & configurations
- Need to be consistent across related configurations
- Need to be intuitive for annotators (easy to remember)

$\rightsquigarrow$ Long-term work: often 1 year to build, then years to improve!

- Conveying ideas: with principles + with examples (prototypes, minimal pairs, edge cases...)

Questions to address: for NER, Mention detection...

- List and describe entity types
- Delineate the exact scope per type, formulate decision criteria (e.g. when is "*France*" a LOC or an ORG)
- Nested or flat (keeping which entity)? Discontinuous entities? Only named entities? Which PoS tags (only PROPNs, or also PRONs...)?
- Guidance on span boundaries: how long should the mention be, should it include determiners, modifiers...?
- ...

Questions to address: for EL

- What entities to consider?
- What mentions to consider?
- What tolerance for near-identity?
- Treatment of unlinkable mentions
- ...

What should Entity Linking link?

Henry Rosales-Méndez, Barbara Poblete and Aidan Hogan

Millenium Institute for Foundational Research on Data
Department of Computer Science, University of Chile
{hrosales,bpoblete,ahogan}@dcc.uchile.cl

Questions to address: for Open IE

- How implicit can a relation be?
- How redundant should triples be (e.g. one specific and one generic)?
- How to choose argument boundaries?
- How to word implicit relations?
- ...

WiRe57 annotation guidelines

3.6 Possessives

Possessives are a special case of inferred relations where the relation is [has].

Mary's dog is brown.	(Mary's dog, is, brown) (Mary, [has], [a] dog)
The prefecture of this city	(this city, [has], [a] prefecture)
The GOP is American. *Its leaders include Ronald Reagan.*	(The GOP, is, American) (Its/(The GOP's) leaders, include, Ronald Reagan) (The GOP, [has], leaders)

3.10 Tuples calling for a single argument

Some relationships have a single argument, typically the subject of the verb, which results in tuples of the form (arg_1, rel). In the example below, grow in popularity is a non-compositional phrase that cannot be expressed as (X, grow in, popularity) and therefore does not call for any argument.

Sangster grew in popularity.	(Sangster, grew in popularity)

3.14 The limits of inference

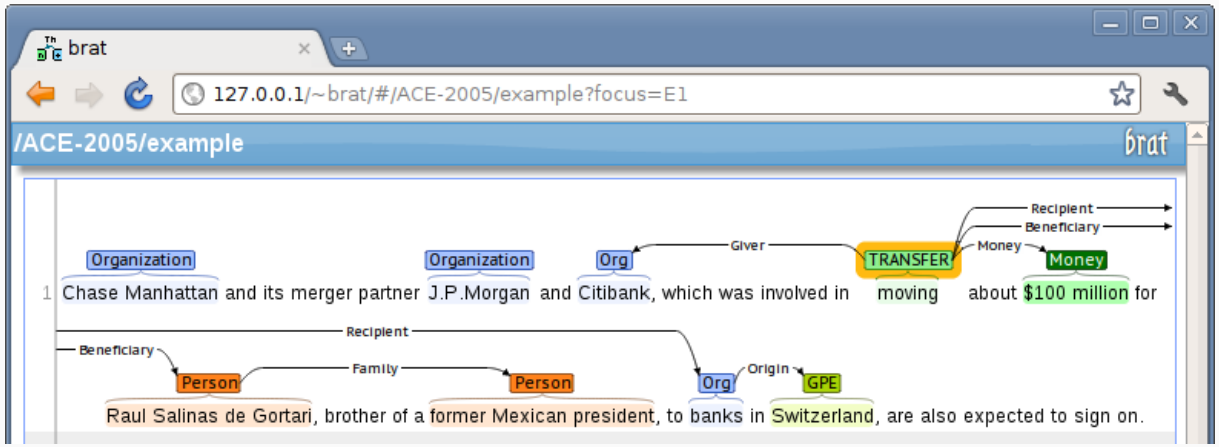
Because the concept of "light inference" is subjective, we propose a few examples and counterexamples that delineate the limits between the two classes.

Jason Charles Beck, a Jewish Canadian musician, was born in 1972.	(Jason Charles Beck, [is], Jewish)
Gonzales is the son of Ashkenazi Jews who were forced to flee from Hungary during World War II.	(Gonzales, [is], Jewish) Complex inference based on culture and human heredity

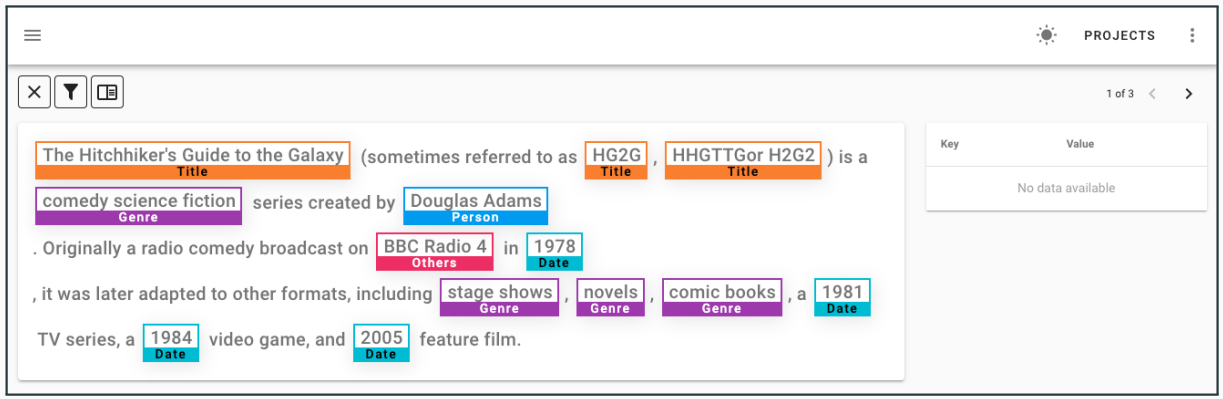
The annotation process

Annotation tools

- brat



- doccano



- More complex tools with automatic detection of inconsistencies: e.g. QA4IE (Jimenez Silva et al., 2022)
- Active learning: the model itself selects the (hardest) sentences to annotate

The learning curve of annotators

- Guidelines are often hard to grasp until you have tried to apply them
- Annotators need to acquire reflexes, know what patterns to be careful of, get used to the annotation tool... a training per se
- The first portion (e.g. 10%) of annotations is always discarded (reannotated at the end)
- It is not wasted: this is part of the process!

Multiple annotators

- Annotators have their own biases, different understandings of the guidelines...
  $\hookrightarrow$ Better annotations if spread over multiple annotators

- If budget allows: each annotator annotates the whole corpus
  $\hookrightarrow$ Adjudication: comparing each set of annotations, discussing divergences, and choosing or merging them into a shared annotation (+ potential update of guidelines)

- If not: assign different parts of the corpus to each annotator + have them all annotate a given part
- If really not: single annotator + another individual annotating just a small overlap for quality control

Inter-annotator agreement (IAA)

- Computed on overlapping annotations: how often do annotators agree, and is it more than expected based on their natural biases? $\rightsquigarrow$ chance-corrected agreement
- Cohen's $\kappa = \frac{\text{Agreement}_{\text{observed}} - \text{Agreement}_{\text{expected}}}{1 - \text{Agreement}_{\text{expected}}}$

	A	B	C	$\sum$
A	$N_{A,A}$	$N_{A,B}$	$N_{A,C}$	$N_{A,*}$
B	$N_{B,A}$	$N_{B,B}$	$N_{B,C}$	$N_{B,*}$
C	$N_{C,A}$	$N_{C,B}$	$N_{C,C}$	$N_{C,*}$
$\sum$	$N_{*,A}$	$N_{*,B}$	$N_{*,C}$	N

Inter-annotator agreement (IAA)

- Computed on overlapping annotations: how often do annotators agree, and is it more than expected based on their natural biases? $\rightsquigarrow$ chance-corrected agreement
- Cohen's $\kappa = \frac{\text{Agreement}_{\text{observed}} - \text{Agreement}_{\text{expected}}}{1 - \text{Agreement}_{\text{expected}}}$

	A	B	C	$\sum$
A	$N_{A,A}$	$N_{A,B}$	$N_{A,C}$	$N_{A,*}$
B	$N_{B,A}$	$N_{B,B}$	$N_{B,C}$	$N_{B,*}$
C	$N_{C,A}$	$N_{C,B}$	$N_{C,C}$	$N_{C,*}$
$\sum$	$N_{*,A}$	$N_{*,B}$	$N_{*,C}$	N

$$A_o = \frac{N_{A,A} + N_{B,B} + N_{C,C}}{N}$$

$$A_e = \frac{N_{A,*} \times N_{*,A} + N_{B,*} \times N_{*,B} + N_{C,*} \times N_{*,C}}{N * N}$$

Inter-annotator agreement (IAA)

- Computed on overlapping annotations: how often do annotators agree, and is it more than expected based on their natural biases? $\rightsquigarrow$ chance-corrected agreement
- Cohen's $\kappa = \frac{\text{Agreement}_{\text{observed}} - \text{Agreement}_{\text{expected}}}{1 - \text{Agreement}_{\text{expected}}}$

	A	B	C	$\sum$
A	$N_{A,A}$	$N_{A,B}$	$N_{A,C}$	$N_{A,*}$
B	$N_{B,A}$	$N_{B,B}$	$N_{B,C}$	$N_{B,*}$
C	$N_{C,A}$	$N_{C,B}$	$N_{C,C}$	$N_{C,*}$
$\sum$	$N_{*,A}$	$N_{*,B}$	$N_{*,C}$	N

$$A_o = \frac{N_{A,A} + N_{B,B} + N_{C,C}}{N}$$
$$A_e = \frac{N_{A,*} \times N_{*,A} + N_{B,*} \times N_{*,B} + N_{C,*} \times N_{*,C}}{N * N}$$

- Krippendorff's $\alpha = 1 - \frac{\text{Disagreement}_{\text{observed}}}{\text{Disagreement}_{\text{expected}}}$, where disagreement is quantified by a distance among annotations
- Low IAA ($< 60\%$) can be OK: some tasks are just too subjective, ambiguity can be real
- More reading: (Artstein & Poesio, 2008)

IAA beyond classification

- Not straightforward! Not completely solved...
- Krippendorff's α can be generalized, using an appropriate distance function: see (Skjærholt, 2014), (Braylan et al., 2022)
- Often, approximated with F1 or similar metrics (ensuring equivalent roles for predictions and references)
- When more than 2 annotators: can approximate by averaging pairwise agreements

Intra-annotator agreement

- Annotators are not machines: it is normal to deviate
- Annotators get tired
- Annotators get influenced by previous sentences
- Annotators evolve in their understanding of the guidelines

↪ Same computation but on the annotator's own (repeated) annotations

Good practices for corpus annotation

Semi-automatic annotation & cognitive biases

- Usually much faster to fix automatically generated annotations, rather than creating them from scratch
  $\hookrightarrow$ Start with a heuristic or a weak model to speed up annotation

- But automation bias: humans over-rely on automated suggestions
  $\hookrightarrow$ All annotators will resolve ambiguities the same way (the machine's), complex cases will not be properly analyzed (defaulting to the suggestion)

- Always a trade-off, make some measurements to assess the risk
- Extra care when using as test data: unfair to evaluate a system with data pre-annotated by the same system

Benefits of disagreement

- If ambiguity is real, disagreement is legitimate

  $\rightsquigarrow$ do not penalize in test?

- For complex sentences, disagreement is an information

  $\rightsquigarrow$ useful for training?

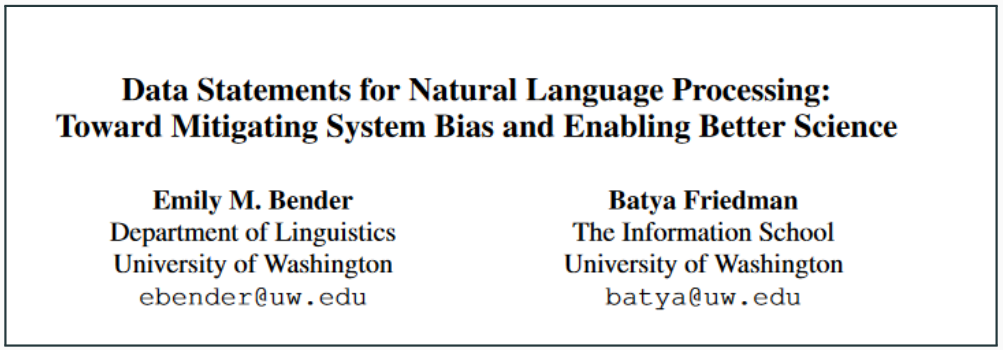
$\hookrightarrow$ Is adjudication always appropriate?

We Need to Consider Disagreement in Evaluation

Valerio Basile[*♣], Michael Fell[♣], Tommaso Fornaciari[♦], Dirk Hovy[♦], Silviu Paun[♥], Barbara Plank[♠], Massimo Poesio[♥], Alexandra Uma[♥]
[♣]University of Turin, [♦]Bocconi University
[♥]Queen Mary University of London, [♠]IT University of Copenhagen
[♣]{valerio.basile, michaelkurt.fell}@unito.it
[♦]{dirk.hovy, fornaciari.tommaso}@unibocconi.it
[♥]{s.paun, m.poesio, a.n.uma}@qmul.ac.uk, [♠]bplank@itu.dk

Documenting a corpus

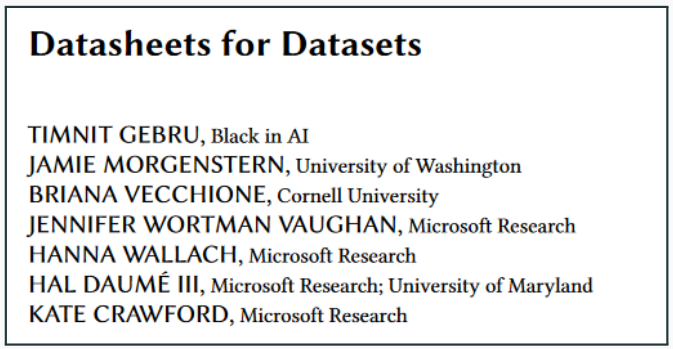
- Source of raw data (URL, dates...), any preprocessing, filtering
- Language, dialect, variety...
- Speaker information (age, gender, socioeconomic status...)
- Annotator information (age, gender, socioeconomic status, background...)
- Format, annotation scheme, annotation process & quality
- License, contacts, version number
- ...



Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science

Emily M. Bender
Department of Linguistics
University of Washington
ebender@uw.edu

Batya Friedman
The Information School
University of Washington
batya@uw.edu



Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

Crowd-sourcing & Amazon Mechanical Turk

Technical challenges:

- Do the annotators have the necessary skills?
  $\rightsquigarrow$ design proficiency tests
- Are they doing the task seriously enough?
  $\rightsquigarrow$ regular checks with repeated inputs & tests
- Higher risk to get lazy
  $\rightsquigarrow$ monitor intra-annotator agreement carefully
- Less contact means less guidance: more efforts on the guidelines
- Annotations spread over many annotators: need more redundancy, more complex adjudication

Ethical concerns:

- Pay? Working conditions?
- Often asked to report on those nowadays

$\hookrightarrow$ Strict policy in some labs against using Mechanical Turk

Legal considerations

- Before distributing data: check that you are allowed to
- Before using the data: same!
- Having access to data does not mean being allowed to use it (nor to redistribute it): Internet contents are not free to use
- What is the licence of the original texts? What is the licence of your annotations?
- If it contains personal data (even without names it can be personal): check compliance with the regulation (GDPR)
- Intellectual property, authors' rights, owner rights, copyright...

$\rightsquigarrow$ If you're unsure: **ask for counsel**. Labs and companies usually have a legal expert to assist (+ data protection officer).

Maintaining and versioning

- Errors will likely remain in your annotations: think of how you will update them when you or others detect errors
- Annotations can also be revised if you update your guidelines
- Keep track of all changes applied: git can help, see e.g. `https://github.com/universaldependencies`
- Ensure reproducibility of experiments: use version numbers to reference different versions of the annotations (& the data)
- If releasing your corpus publicly: get an ISLRN, register in the LRE Map...

The project

Overview

- Experience the annotation process yourself: for Open IE, starting with existing guidelines (WiRe57)
- Use that corpus to evaluate open source tools for Open IE
- Group project (3 students, or max 4 if necessary): annotate individually then adjudicate
- Main job is to explain what you did and to think about it, not to annotate "perfectly"

Data & guidelines

- Corpus of 40 sentences (already tokenized): online here
- Follow the WiRe57 annotation guidelines (online here)
- Reading the WiRe57 paper (Léchelle et al., 2019) will help for the methodology and for understanding the guidelines
- Annotation format (plain text):

```
#1: Lorem dolor sit amet .
Lorem <TAB> sit amet <TAB> dolor
#2: Sed non risus .
...
```

Annotation process for the project

1. Each student annotates sentences 1-10 alone (manual, fully individual, no discussion with others or you will be biased)
2. Group discussion to share your views and debate disagreements, decide on refined guidelines
3. Each student annotates sentences 11-20, using the group's refined guidelines
4. Group discussion for feedback and to finalize the guidelines
5. Each student annotates sentences 21-40, then reannotates from scratch (no peeking, wait a bit to forget) sentences 1-10
6. Group adjudication: compare all versions for each sentence and decide on a common annotation

Annotation process for the project – quality control

Compute:

- Intra-annotator agreement between first and second versions of sentences 1-10
- Inter-annotator agreement on the first version of sentences 1-10, then on the second (compare)
- Inter-annotator agreement on the complete corpus (1-10 v2 + 11-20 + 21-40)
- Agreement of each annotator with the adjudicated annotation

Evaluation of Open IE tools

- Consider the adjudicated corpus as test data
- Evaluate 2 open source tools for Open IE in English: at least Stanford Open IE + another of your choice (see suggestions in the WiRe57 paper)
- Use 2 metrics: at least the F1 as applied in the WiRe57 paper + another variant of F1 (e.g. with inferred words, with a different matching of predicted/references, with exact or partial match of arguments instead of token-weighted...)
- Compare both tools, compare both metrics, and comment
- Write your own implementation of the metrics (not an existing script)
- Data formats won't match, some conversion code will be needed: it is part of the assignment, don't do it manually

First submission

By December 30 (23:59): individual submission of

- annotations for sentences 1-10 (v1)
- short report (1-2 pages) explaining your choices, impressions, concerns, ideas...

↪ By email to me, titled "[IE-AKGC] Individual report for FirstName LastName" + attachments

The earlier you start, the easier it will be: annotation work is difficult to rush.

Second submission

By January 15 (23:59): group submission of

- Each student's full annotations (separate files for sentences 1-10 v1, 1-10 v2, 11-20 and 21-40) + the adjudicated ones
- Report (7-8 pages):
  - ∼3 p. relating your disagreements, the choices you made as a group and why, and reporting and commenting your IAAs
  - 1-2 p. proposing ideas and suggestions (e.g. complements or refinements to the WiRe57 annotation guidelines)
  - ∼2 p. about the evaluation results and your analysis and comments
  - ∼1 p. on your impressions on how easy it would be to automatically extract a knowledge graph from this text, the challenges you see
- Code for computing the IAAs on your data (when run as-is from your folder, it outputs the same numbers as the report's)
- Code for computing the evaluation metrics on your data

↪ By email to me (the whole group in cc), titled "[IE-AKGC] Group report for First1 Last1 + First2 Last2 + First3 Last3" + single attachment (zip)

Grading criteria

- Richness and depth of thinking when annotating
- Compliance with the annotation process

  $\hookrightarrow$ Cheating to get perfect agreement will lower your grade
- Rigor in evaluation
- Quality of analysis and hindsight (including suggestions for better guidelines)

$\rightsquigarrow$ Part of the grade will be individual, part of it common to the group

- If late: 2-day tolerance with penalty points, then zero
- Results: early February

Licensing

- I won't: reuse or distribute your annotations (except if you explicitly tell me to)
- I might: draw inspiration from your feedback, as an input to future research
- If you want to distribute your annotations: you need permission of the whole group, they are the result of team work
- Raw texts are CC BY-SA 4.0 (EN Wikipedia page on Noam Chomsky)

See you next week!

first.last@inria.fr