DFKI Annotation Guidelines

Version: 1.0
Author(s): Saskia Schön, Sebastian Krause, Veselina Mironova, Leonhard Hennig, Aleksandra Gabryszak, Philippe Thomas

Scope of this document

The annotation guidelines in this document are based on the

- ACE annotation guidelines (English entities & relations): https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications
- Timex annotation guidelines: http://www.timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf

The main difference to ACE guidelines is the treatment of GPE entities - we prefer LOC or ORG, in particular for countries, cities, regions, federal states, counties etc. Only if the context implies the entity is used with a focus on the the government aspect, use GPE.

For date and time expressions, when in doubt, refer to the Timex annotation guidelines, or try using Stanford's http://corenlp.stanford.edu/ online demo (at least for English, it mostly works correctly).

TODO: merge contents of Business Corpus Guidelines (https://docs.google.com/document/d/1xwwMcNL6Rpy-H1SpIZM0pdpcNLrwBlDN7vza6B5pY1M/edit)

1 Entities

Mention Extent

In general, token boundaries (whitespace & co) should determine the mention extent. "-" or "/" typically count as a token boundaries (for exceptions, see below). For tweets, # and @ should not be included in the mention extent (unless they occur inside a multi-token entity, e.g. "#[Flughafen_#Tempelhof]").

The mention extent should be as long as "needed" to denote a specific entity. It may include adjectives, numerals ("more than", "a few", "some", "several"), numbers etc., if these are used to denote a specific subset of a set-based named entity mention (e.g. plural forms, see examples below).

If an entity is referred to by 2 subsequent token (sequences), e.g. "company (acronym)", "street-number street-name", or "route-identifier start-loc - end-loc", annotate 2 separate entities. For locations, this is typically mostly relevant for "americanized" city names, e.g.

“New York, NY” should be tagged as [New York] = location-city, [NY] = location, and NOT “[New York, NY] = location-city”. The rule also holds for date-time expressions, e.g. “Tue, March 21st, 2011, 6:30 p.m.” resolves to [Tue, March 21st, 2011] = date, [6:30 p.m.] = time.

As a rule, unless otherwise specified for a particular annotation task, do not annotate nested mentions unless required for a relation mention, e.g. in “PD Zwickau ...”, annotate “PD Zwickau” as an organization (police department zwickau), but do not annotate a 2nd entity “Zwickau” as a location-city.

Examples

- “Zuffenhausen-Süd (Friedrichswahl) in Stuttgart-Feuerbach “ in “B27 Stuttgart Richtung Mühlacker zwischen Zuffenhausen-Süd (Friedrichswahl) in Stuttgart-Feuerbach und Korntal Baustelle, linker Fahrstreifen gesperrt, 2 km Stau”
    - [Zuffenhausen-Süd] = location
    - [Friedrichswahl] = location

Do not include cut off words in the mention extent:

Examples

- smartdata/tweet_470.xml: RT @DB_Info: Ersatzverkehr auf der Linie RE 4 zwischen Torgelow und Ueckermünde Stadthafen vom 2. September bis 12. D…
    - [12. D] is not the correct extent for the last date entity, rather use only [12.]

Tokenization exceptions:
“/” does not always constitute a token boundary:

Examples

- smartdata/news_928.xml: “Polizei Wertheim/Main-Tauber … Mit einer Betrugsmasche, …”,
    - annotate a single organization “[Polizei Wertheim/Main-Tauber]”.

For the remainder of the document, square brackets [ ] will indicate the extent of an Entity Mention.

Organization

An Organization entity must have some formally established association. Typical examples are businesses, government units, sports teams, and formally organized music groups. Include acronyms.

See Section 3.2 in
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

Examples

- smartdata/tweet_462.xml Borussia Mönchengladbach gegen den FC Sevilla im Borussia Park #BMG #Sev #BMGvSEV #BMGSEV @SevillaFC_ENG @Borussia
    - [Borussia Mönchengladbach] = organization
    - [FC Sevilla] = organization
    - any occurrence of [BMG] and [SEV]/[Sev], even in the last hashtag = organization
- smartdata/tweet_1147.xml: RT @SZ: #EU und #Türkei einigen sich auf Flüchtlingsabkommen
    - [EU] -> organization
- smartdata/tweet_753.xml: RT @PolizeiMuenchen: Wir hätten wohl tatsächlich…
    - [PolizeiMuenchen] -> organization
- smartdata/tweet_460.xml: PD Zwickau – Zwickau: Radfahrerin bei Verkehrsunfall schwer verletzt https://t.co/Y4RwGtwNVL
    - [PD Zwickau] = organization

Organization-Company

Subtype for companies.
Public transport “organizations” are considered as companies, e.g. “BVG”, “Münchner SBahn”, etc. Similarly, online news sites (“faznet”, “rpo”, “ntv.de”, “rbbonline”, “wdr.de”) are companies.

Examples

- smartdata/tweet_410.xml: Plant Amazon die Übernahme des Flughafen Frankfurt Hahn? https://t.co/r82LH1H4MA https://t.co/iRWoWvfjiw
    - “Flughafen Frankfurt Hahn” = organization-company
- smartdata/tweet_224.xml Google-Ableger will Staus mit Big Data bekämpfen: Die zum Google-Imperium gehörende Firma Sidewalk Labs hat si... https://t.co/sm6J0jBYtL
    - “Google-Ableger” = organization-company
- smartdata/tweet_521.xml: Münchner S-Bahn: SPD regt sich über Dobrindt auf: Totengräber der 2. Stammstrecke https://t.co/ZskkajmYCP #Abendzeitung
    - “Münchner S-Bahn” = organization-company (although, in this example, one could argue that the term does not quite exactly refer to the company MVV, more to the S-Bahn in general?)
    - “Abendzeitung” = organization-company
- smartdata/tweet_519.xml: @StN_News Um 8.30 war das Maximum der Verspätungen noch nicht überschritten, wie diese Diagramme belegen https://t.co/vf44SY1GCk #SBahnStgt
    - “StN_News” = organization-company
    - “SBahnStgt” = organization-company
- smartdata/tweet_509.xml: Streik bei Londons U-Bahn gegen 24-Stunden-Betrieb: Aus Protest

- “Londons U-Bahn” = organization-company

Person

See Section 3.1 of
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

Location

Places defined on a geographical or astronomical basis which are mentioned in a document ~~(and do not constitute a political entity)~~ give rise to Location entities. These include, for example, the solar system, Mars, the Hudson River, Mt. Everest, and Death Valley. Places distinguished only by the occurrence of an event at that position ("the scene of the murder", "the site of the rock et launching") are not entities.

Important: Contrary to ACE guidelines, include nationalities/countries (“the Turkish police”) as Location instead of GPE, similar for cities, counties, federal states, continents etc.
Important: Please also read Sec 3.4.3 “Non-Locations” in the ACE guidelines very carefully to avoid tagging non-locations as locations!

Use sub-types (see below) if applicable! The only exception are exits on highways / autobahnen, where we use “location” as the NE type, even if they might refer to a city. For (train) routes, always use location-stop for route endpoints. Also, do not use a subtype if the NE stands for a larger entity, e.g. “Berlin” for the government/country (-> GPE). Terms such as “beide Richtungen”, “je Richtung”, Richtung “Norden” or “stadteinwärts” are also locations. For traffic-related locations, include specifiers when required, e.g. “Kreuz”, “Dreieck”, “Anschlussstelle”, “AS”, “Abzweig (nach)”, “Verzw.” (Schweiz) etc in the location extent. However, do not include “Ecke” oder “Kreuzung” in city street specs, e.g. in “[Unfall] bei [Warschauer Str.] Ecke [Frankfurter Allee]”.
Also include “Kreis” or similar terms if it is necessary to differentiate the location from the city, e.g. “[Kreis Tuttlingen]: Unfall … “.

Examples

- smartdata/rss_8.xml: A99 Ostumfahrung München Richtung Nürnberg zwischen Kreuz München-Ost und Kirchheim Behinderungen nach einem LKW-Unfall
  - “Kirchheim” = location
- smartdata/tweet_430.xml: RT @faznet: Unabhängig von Madrid? #Katalonien beschließt die Abspaltung von Spanien
  - “Madrid” = location (actually, GPE), not location-city
- smartdata/tweet_429.xml: RT @SPIEGELONLINE: #Katalonien hat eine Resolution zur Abspaltung von Spanien verabschiedet. Bis 2017 will die Region unabhängig sein.
  - “Region” = location

- smartdata/tweet_747.xml: RT @SPIEGELONLINE: Frankreichs Premier Valls hat die EU aufgefordert, ihre Grenzen für Flüchtlinge aus dem Nahen Osten zu schließen
    - “Nahen Osten” = location
- smartdata/tweet_319.xml: Vollsperrung der #A100 in Fahrtrichtung Nord, Grund: schwerer Unfall
    - “Nord” = location
- smartdata/tweet_886.xml: #Hamburg: Die Bahrenfelder Chaussee ist stadteinwärts zwischen Theodorstraße und Von-Sauer-Straße bis Ende Juli gesperrt.
    - “stadteinwärts” = location
- smartdata/tweet_716.xml: Auf den Linien S5, S51 und S52 kommt es in beiden Richtungen zu Verspätungen.
    - “beiden Richtungen” = location
- smartdata/rss_230.xml: Kreis Tuttlingen, K5900 zwischen Abzwe ….
    - “Kreis Tuttlingen” = location

If a location is used to refer to some other entity type, do not annotate as location:

Examples

- smartdata/news_20.xml: “Londoner Börse legt Jahreszahlen vor: Deutsche wollen Fusion”
    - “Deutsche” -> organization-company (from the context of the doc, which clarifies that “Deutsche” is coreferent with “Deutsche Börse AG”)

Location-City

Annotate city or town/village names as location-city, UNLESS they occur in railway/flight connections (where they are typically considered as referring to a location-stop type of entity)

Highway / Autobahn - use location-city for direction, and general route start-end locations, but use “location” for exits.
Railway / Public transport - use location-city for direction, and location-stop for everything else
Flight / Airport - use location-city in general, location-stop if the reference gives the specific airport used (“Heathrow”, “MUC”).

Examples

- ■ #Hamburg: Grusonstraße - Wöhlerstraße zwischen Bredowbrücke und Liebigstraße ist wegen #Bauarbeiten bis zum 23. Oktober …
    - “Hamburg” = location-city
- smartdata/tweet_1282.xml: A61 Koblenz Richtung Mönchengladbach zwischen Swisttal und Kreuz Bliesheim 2 km Stau
    - “Koblenz” = location-city
    - “Mönchengladbach” = location-city
    - “Swisttal” = location
    - “Kreuz Bliesheim” = location
-

NOT a location-city, but location-stop

- smartdata/tweet_469.xml: [DB Regio] 1. Akt. #Günzburg - #Mindelheim: #Störung an einem #Bahnübergang / #Schienenersatzverkehr
    - “Günzburg” = location-stop
    - “Mindelheim” = location-stop
- smartdata/news_876.xml: Swiss-Flugzeuge von und nach München bleiben am Mittwoch am Boden.
    - “München” = location-stop (see Section on Metonymy)

Location-Street

Annotating multiple, separate entities should also be adhered to in the case of streets having a name and a number, e.g.

- smartdata/tweet_500.xml Hat Fahrern in der Nähe geholfen, mit der Meldung ein schwerer stau auf B4R - Von-der-Tann-Straße, Nürnberg mit @waze - Fahre Sozial. htt…
    - “B4R” and “Von-der-Tann-Straße” are 2 location-street entries. A single TrafficJam relation with those 2 streets as location arguments can be annotated

Location-Route

For location-routes, include abbreviation / indicators such as U (for U-Bahn) or A (Autobahn), but do not include generic terms like “Linie”. I.e. include the route type (“U”), which may be needed for disambiguation, but not the trigger term indicating that it is a route (“Linie” or “Li”). Note that “locationA - locationB” may denote a route (or with any other separator besides ‘-’).

Examples

- smartdata/tweet_553.xml: SPD erfreut über Verlängerung der Linie 3 …
    - “3” = location-route, NOT “Linie 3”
- smartdata/tweet_469.xml: [DB Regio] 1. Akt. #Günzburg - #Mindelheim: #Störung an einem #Bahnübergang / #Schienenersatzverkehr
    - “Günzburg - #Mindelheim” is location-route
- smartdata/news_879.xml: Die kurzfristig angekündigte Sperrung der [ICE-Schnellfahrstrecke Hannover-Kassel] in drei Wochen soll nach
    - “ICE-Schnellfahrstrecke Hannover-Kassel” = location-route

Annotate 2 separate entity mentions, as usual, when a route is specified both by a number/identifier and a start/end location (however, adding the 2nd route annotation is not “required”, usually for purposes of relation mention annotation, if you have a route specifier like RE 3, that is enough):

- smartdata/tweet_468.xml: RT @DB_Info: Ausfälle und Ersatzverkehr auf der Linie RE 3 Stralsund/Schwedt (Oder) – Berlin – Elsterwerda vom 10. bis …
    - “RE 3” = location-route
    - “Stralsund/Schwedt (Oder) - Berlin - Elsterwerda” = location-route (optional)

Note that even “isolated” numbers may indicate a location-route, e.g. in “table”-like tweets:

- smartdata/tweet_711.xml: 170: Hohes Verkehrsaufkommen Verspätungen von bis zu 20 Minuten....
    - From the context, it is likely that “170” refers to a location-route (e.g. bus route), and thus there is a “Delay” relation mentioned in this tweet

Location-Stop

Entities mentioned in the context of public transport events are often location-stops, even if they are just city names (See Section on Metonymy!).
Airport names/codes are also location-stops, not locations (unless the airport name is used to refer to its operating company, e.g. “Amazon is planning to take over the airport Frankfurt Hahn”)

Examples

- smartdata/tweet_470.xml: RT @DB_Info: Ersatzverkehr auf der Linie RE 4 zwischen Torgelow und Ueckermünde Stadthafen vom 2. September bis 12. D…
    - “Torgelow” -> location-stop
- smartdata/tweet_469.xml: [DB Regio] 1. Akt. #Günzburg - #Mindelheim: #Störung an einem #Bahnübergang / #Schienenersatzverkehr
    - “Günzburg” and “Mindelheim” are location-stops
- smartdata/tweet_478.xml: RT @SPIEGEL_Reise: Rinjani macht Ärger: Flughafen auf Bali wegen Vulkanausbruch gesperrt…
    - “Flughafen auf Bali” = location-stop

Use location-city, not stop:

- smartdata/news_875.xml: Ich hatte mal eine Überholung im RE 7 in Holzwickede Richtung Hamm,
    - “Hamm” = location-city

Disaster-Type

Includes natural and man-made disasters as specified in
http://www.ifrc.org/en/what-we-do/disaster-management/about-disasters/definition-of-hazard/

Only annotate if not used as a metaphor, e.g. “flood” and “earthquake” should NOT be annotated in texts relating eg. “landslide election victories” or “flood of immigrants” etc.

Trigger

Annotate triggers ONLY if a relation is expressed in the sentence (ie. if you annotate a relation, also annotate triggers, if available). Triggers can be single words or phrases, or even “split” phrases, e.g. in German “[legten] die Angestellten [die Arbeit nieder]”. Similar to disaster-type, do not annotate a trigger if it is used metaphorically or not in the sense of the relation to be annotated, e.g. “landslide election victories” or “flood of immigrants”, or “strikes” in “This strikes me as unusual”.

For trigger entities, you do not need to always include “specifying” appositions in the extent:

Examples

- smartdata/tweet_316.xml: Nach Mitternacht kam es bei Zügen der Linie S2 zu größeren Verspätungen wegen einer technischen Störung am RE 19463 nach Aalen #SBahnStgt
    - marking “technischen Störung” as the trigger mention is sufficient, using “technischen Störung am RE 19463 nach Aalen” would be unnecessarily specific
- smartdata/tweet_312.xml: RT @earlybird445: Wegen Weichenstörung in #Zuffenhausen kommt es bei S4, S5 und S6/S60 zu großen Verspätungen und Ausfällen. Keine Infos be…
    - marking “Weichenstörung” is sufficient. “Weichenstörung in #Zuffenhausen” as a trigger for the delays/cancellations of routes S4, S5 etc, is too specific

Examples when NOT to annotate a trigger:

- “Chinesen auf Einkaufstour in Europa”
    - do not annotate “Einkaufstour” for a possible Acquisition relation, since no concrete/specific event is actually mentioned
- “Mein Mann streikt schon wieder”
    - “streikt” ist not a trigger since it is used metaphorically / not in the sense of the “strike” relation we want to extract

Date and Time expressions in general

When in doubt, refer to the Timex annotation guidelines (http://www.timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf) , or try using Stanford’s http://corenlp.stanford.edu/ online demo (at least for English, it mostly works correctly). Prepositions like “on”, “to”, “at”, “until” are typically not included in the extent of dates / times / durations.

Date

Specific dates, e.g. “March 21st, 2012”, “March, 21st”, “2011”, “early October”. Holds also for dates that “appear” like a time span, e.g. “since 2011” or “until Christmas”, and includes day-of-week info. Include ‘.’ in dates, e.g. “am 7. und am 8. 12.” -> “am [7.] und am [8. 12.]”.

Choose “date” if it answers a “When” question.

Do not annotate a date if terms are used in a non-date context, e.g. as a song title (“Yesterday”, “In the year 2525”) or in radio claims such as “Das Gute von gestern mit dem Besten von heute”.

See Section 2.2.3 of http://www.timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf :
DATE: The expression describes a calendar time.
Mr. Smith left [Friday, October 1, 1999]
[the second of December]
[yesterday]

in [October of 1963]
in [the summer of 1964]
on [Tuesday 18th]
in [November 1943]
[this year's summer]
[two weeks from next Tuesday]
[last week]
in [3 years]
in [CW 11]

(do not include the italic prepositions in the mention extent)

Examples

- smartdata/tweet_551.xml: Die 24 Türchen bis Weihnachten sind für viele Eltern der Horror.
    - "Weihnachten" -> date
- smartdata/tweet_271.xml: RT @faznet: Historische Dürre - die Sierra Nevada ist so schneearm wie seit Kolumbus nicht mehr. http://t.co/fCq7xhinnB
    - seit "Kolumbus" -> date
- smartdata/tweet_1324.xml: 2 #Flugzeuge im Umkreis #Flughafen #Hannover sichtbar via #ADSB am Mon Jul 18 2016 05:35:26 GMT+0200 (CEST) mit #Raspberry Pi
    - "Mon Jul 18 2016" -> date
- smartdata/tweet_1310_future.xml: Piloten von Air Europa wollen zur Hochsaison in den Streik treten: …
    - "Hochsaison" -> date
- "Dienstag, 22. März, 8.00 – 18.00"
    - [Dienstag, 22. März] = date
    - [8.00] = time
    - [18.00] = time
- smartdata/rss_54.xml: "EN 446 nach Köln Hbf (17.48 Uhr ab Warszawa Wschódnia bzw. 23.33 Uhr ab Berlin Ostbahnhof) wird in der Nacht 13./14. März von Berlin Hbf bis Hannover Hbf umgeleitet und hält nicht in Potsdam Hbf, Brandenburg Hbf, Magdeburg Hbf und Braunschweig Hbf."
    - "Nacht" = time
    - "13." = date
    - "14. März" = date
- smartdata/news_898.xml: "KW 11: Analysten-Flops der Woche…"
    - "KW 11" = date

Time

Non-repeating, non-spanning time expressions, e.g. "8:30 p.m.", etc. Include "Uhr" / "o'clock" in the mention extent, as well as timezone information ("GMT+0200"). Time-related terms like "Betriebsschluss", "abends", etc, are also to be marked up as "time".

See Section 2.2.3 of http://www.timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf :
TIME: The expression refers to a time of the day, even if in a very indefinite way (as in the two last examples below):

Mr. Smith left [ten minutes to three]
at [five to eight]
at [twenty after twelve]
at [half past noon]
at [eleven in the morning]
[late last night]
bis [5 Uhr früh]

Contrary to TimeML, we do not consider the examples below as time, but rather as time + date
[the morning]time of [January 31]date
at [9 a.m.]time [Friday, October 1, 1999]date
[last [night]time]date ??

Examples

- smartdata/tweet_1130.xml: 7 #Flugzeuge im Umkreis #Flughafen #Hannover sichtbar - via #ADSB am Thu Jun 23 2016 00:18:35 GMT+0200 (CEST) mit #Raspberry Pi
  - "00:18:35 GMT+0200" -> time (without "CEST")
- smartdata/rss_251.xml: jeweils 5.00 Uhr – Betriebsschluss …
  - "Betriebsschluss" = time
  - "5.00 Uhr" = time

Duration

Time spans with a unit, e.g. "2 hours", or "the [18-year-old] teenager", or "ganztägig". The unit may be implicit in cases like "in 4 to 6 weeks" -> "4 to 6 weeks" is also a "duration". Used also in constructions with "for", e.g. "we prepared for this event for months". Numeral-like words should be included in the mention extent, eg. "several months". When in doubt, use date instead of duration, eg. for "in 3 years". Include "." if unit is abbreviated (reason: tokenization will typically not separate the '.' from the previous token, plus it is more consistent with other annotations, e.g. in titles/honorifics etc.)
See Section 2.2.3 of http://www.timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf :

Choose "duration" if it answers a "How long" question

DURATION: The expression describes a duration. This value is assigned only to explicit durations like the following:

Mr. Smith stayed [2 months] *in Boston*
[48 hours]
[three weeks]
[all last night]
[20 days] *in July*
[3 hours] *last Monday*.

[2 - 3 hours].
[eine Viertelstunde]
[20 Minuten]
[mehr als eine halbe Stunde] // since “mehr” counts as a numeral

Examples

- smartdata/tweet_… xml: Ein 20-jähriger Mann …
  - “20-jähriger” = duration
- smartdata/tweet_242.xml: RT @rbbabendschau: Flughafengesellschaft bereitete sich offenbar seit Monaten auf #Imtech-Pleite vor: http://t.co/lDjlENFGTW #BER
  - “Monaten” -> duration
- smartdata/tweet_508.xml: @Sassy3009 Aufgrund des hohen Eingangsvolumens durch den Streik, kann es zwischen 4 und 6 Wochen dauern.
  - “zwischen 4 und 6 Wochen” = duration
- smartdata/rss_280.xml: … CNL 471 nach Zürich HB (planmäßig 21.33 Uhr ab Berlin Gesundbrunnen) fährt bis zu 46 Min. früher
  - “46 Min.” = duration (!not “46 Min”)
- smartdata/rss_258.xml: am Sonntag, 15. Mai, ganztägig
  - “ganztägig” = duration

When to use date

- smartdata/tweet_271.xml: RT @faznet: Historische Dürre - die Sierra Nevada ist so schneearm wie seit Kolumbus nicht mehr. http://t.co/fCq7xhinnB
  - “Kolumbus” -> date (arguably very fuzzy …)
- smartdata/tweet_1006.xml: @SPIEGELONLINE #netflix Die sind in 3 Jahren verschwunden oder aufgekauft.
  - “3 Jahren” -> date (one could argue that this is a duration, but the focus of the sentence is on the fact that on a date at latest 3 years from now, Netflix will belong to some other company, not that it will take a time period of 3 years for Netflix to be bought up)

Set (Time-related)

See Section 2.2.3 of http://www.timeml.org/publications/timeMLdocs/annguide_1.2.1.pdf :
SET: The expression describes a set of times. This value is assigned to expressions such as those in section 3.5 of TIDES(02). For example:

John swims [twice a week]
[every 2 days]
on [mondays]

Examples

- smartdata/rss_260.xml: … ICE 777 nach Stuttgart Hbf (planmäßig 8.18 Uhr ab Mannheim Hbf) fällt montags bis freitags von Mannheim Hbf bis Stuttgart Hbf aus. …

    - “montags” = set
    - “freitags” = set

Numbers

Annotate number + unit, if possible, for currency values, as well as for distances and durations (eg. age-related numbers such as “18-year-old”). For distances, as described below, use the “distance” subtype, for time-related numbers “duration”. Do not include units in all other cases, e.g. “6 employees” -> “[6] employees”, etc. Include ordinals (“erster”, “1st”). Do not include ‘.’ in the extent for ordinals (e.g. “1. update” -> “1” is number, not “1.”). For percent values, include the “unit” if it is the %-sign as well as if it is “percent” text.

Examples

- [$ 3.7 million]
- [3 percent]
- [3 %]
- [3.7 million Euro]
- [3-4%]
- [3] to [4 percent]
- smartdata/tweet_439.xml: Deutsche Bahn plant Stellenabbau im Güterverkehr: Die Deutsche Bahn plant zum ersten Mal seit Jahren einen Ste…
    - “ersten” -> number
- smartdata/news_897.xml: Die Gornergrat Bahn konnte den Umsatz um 5,9 Prozent auf 24,7 Millionen Franken verbessern.
    - “5,9” = number
    - “24,7 Millionen Franken” = number

Distance

Number + length unit, especially for traffic jam lengths, should be annotated as “distance“ entities. Not all “number + length unit” are distances, though, e.g. “[8mm] screw” is probably not.

Examples

- smartdata/tweet_895.xml: Ich als Automatik-Verachterin nach 5 km im Stau
    - “5 km” = distance

GPE

Geo-Political Entities are composite entities comprised of a population, a government, a physical location, and a nation (or province, state, county, city , etc.).
See Section 3.3 in
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

Note: Currently, we do not really annotate GPEs. Countries -> Location, things like “EU” -> Organization. We could consider using GPEs for governments etc...

Fictional Named Entities

Do not annotate names that do not refer to a real-world entity (e.g. names “invented” by the author)

Examples

- smartdata/tweet_214.xml: @MeinFernbus was ist denn bei der Fusion damals schief gelaufen, dass ihr jetzt nicht MeinFlixbus heißt? Das stand doch sicher zur Debatte.
    - “MeinFlixbus” contains the real-world entity Flixbus, but the full mention text “MeinFlixbus” is made up and does not refer to a real world entity

Named Entities vs Common Nouns

Commons nouns should be annotated like an entity if they refer to a specific real-world entity or a set of entities via coreference. This also holds if they refer to a real-world location or a time/date reference.
Do not annotate common nouns if the relations mentioned are very “broad-scoped” in time/place, e.g. “Chinesen übernehmen immer mehr deutsche Autozulieferer” - refers to a set of events instead of a single event. (maybe rule - if there is a concrete single event and at least 1 concrete entity mentioned, then annotate, otherwise, do not annotate)

Examples

- RT @jungewelt: Feindliche Übernahme #Fraport übernimmt rentable Flughäfen in #Griechenland https://t.co/jMoGFnfH97 F:Reuters https://t.co/R…
- Fraport übernimmt 14 griechische Flughäfen
    - For an “Acquisition” relation with buyer=Fraport, acquired = "Flughäfen in Griechenland" in ex. 1, resp. “14 griechische Flughäfen” in ex. 2., since “Flughäfen” (indirectly) refers to a set of entities representing companies (otherwise they can’t be acquired in the first place). Similar to “marriage” relation in “the spouses traveled to …”
- smartdata/tweet_505.xml: Flughafen-Streik: Lufthansa streicht fast 900 Flüge - SPIEGEL ONLINE https://t.co/HOfltRTPiC
    - There is a strike relation between “Flughafen” and trigger “Streik”. “Flughafen” corresponds to a set of airport companies, which would have to be resolved via coreference… Note that while the strike affects Lufthansa, Lufthansa is not the company being addressed by the strike.
- smartdata/tweet_416.xml: Der amerikanische Kreditkartenanbieter #Visa kauft für 21,2 Milliarden seine Europa-Tochter zurück. https://t.co/DYsrISg15L #Fusion
    - “Europa-Tochter” = organization-company
- smartdata/tweet_482.xml: ■ #Hamburg: #B431 Stresemannstraße stadteinwärts in Höhe Kaltenkirchener Straße behindern #Bauarbeiten bis zum 31. Juli den …
    - “stadteinwärts” = location
- smartdata/tweet_429.xml: RT @SPIEGELONLINE: #Katalonien hat eine Resolution zur Abspaltung von Spanien verabschiedet. Bis 2017 will die Region unabhängig sein.
    - “Region” = location

- smartdata/tweet_478.xml: RT @SPIEGEL_Reise: Rinjani macht Ärger: Flughafen auf Bali wegen Vulkanausbruch gesperrt…
    - “Flughafen auf Bali” = location-stop
- smartdata/tweet_224.xml Google-Ableger will Staus mit Big Data bekämpfen: Die zum Google-Imperium gehörende Firma Sidewalk Labs hat si... https://t.co/sm6J0jBYtL
    - “Google-Ableger” = organization-company
- smartdata/tweet_1288.xml: RT @SPIEGELONLINE: Lage in der #Türkei: @lufthansa hat alle Flüge umgeleitet, die auf dem Weg in die Türkei waren.
    - “alle Flüge” = location-route
- “Die [EC-Züge] zwischen ... und ... fallen aus”
- “Die [Strecke] zwischen … und … ist gesperrt”
- “auf der [Bahnstrecke Kempten-Immenstadt] der Streckenabschnitt zwischen Immenstadt und Martinszell im Moment gesperrt.”
- smartdata/news_879.xml: Die kurzfristig angekündigte Sperrung der [ICE-Schnellfahrstrecke Hannover-Kassel] in drei Wochen soll nach
    - “ICE-Schnellfahrstrecke Hannover-Kassel” = location-route

Do NOT annotate named entities if their NE type is not in the scope of the current annotation task.

Examples

- smartdata/tweet_949.xml: RT @broeselbub: Wer diesen #EM16-Spielplan geschrieben hat, ist auch in der Terminverwaltung für #BER und #S21 tätig.
    - “S21” is not a location-stop or location-company - in this context, it actually refers to the project of rebuilding the train station.
    - Analogously, “BER” does not refer to the company or location-stop in this context

Nicknames / Metonymy

See the discussion on metonymy in Section 6.2 of the ACE annotation guidelines:
“Metonymy occurs when a speaker uses a reference to one entity to refer to another entity (or entities) related to it. Typical examples are capital city names standing for the country/government, or city names standing for sports teams.” In case of metonymy, annotate the entity type that is ‘indirectly’ refered, e.g. instead of city, use GPE or ORG, etc.

e.g. “New York defeated Boston 6-3” - New York and Boston are ORG, not LOC, because they refer to the sports team.

https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

Location-stops are often denoted by the city name, e.g. in flight or railway connections.

Similarly, for country names used as referents for a company, annotate as organization-company, not location:

- smartdata/news_20.xml: “Londoner Börse legt Jahreszahlen vor: Deutsche wollen Fusion”
    - “Deutsche” -> organization-company
- smartdata/news_20.xml: “Die Deutsche Börse und London Stock Exchange wollen fusionieren. Anleger dürften dabei auch auf mögliche Aussagen zur geplanten Fusion mit der Deutschen Börse achten. Deren Chef Carsten Kengeter hatte auf ein Gegenangebot des US-Rivalen Intercontinental Exchange (ICE) für die Londoner gelassen reagiert.”
    - “Londoner” = organization-company

2 Relations

Annotate only explicit relation mentions that occur within a single sentence with all required arguments. Choose the “closest” arguments, i.e. the relation should have the shortest possible span (e.g., if a clause contains a pronoun and the relation mention, annotate with pronoun as the argument, and add an Identity relation). Annotate future or planned relation mentions, but rename the file by appending a suffix “_future.xml”. Do NOT annotate negated relation mentions, or events marking the end of a relation (“e.g. Traffic jam has dissolved”), but rename the file by appending a suffix “_negation.xml”.

If the relation definition specifies exactly 1 required argument of a type, but the sentence contains a conjunction of multiple entities for this argument, specify multiple relation mentions. Or discuss if we should loosen the relation definition (e.g. for Acquisition, we need exactly 1 buyer argument, but sometimes “multiple” companies act as buyers…)

Examples

- smartdata/tweet_1267.xml: RT @eurobahn_info: #RB67 Schienenersatzverkehr zwischen Münster und Beelen bzw. Rheda-Wiedenbrück vom 12.7. bis 3.8.
    - 1st “RailReplacementService” between Münster and Beelen
    - 2nd “RailReplacementService” between Münster and Rheda-Wiedenbrück

Factual Event or Relation

Examples

- smartdata/tweet_211.xml: RT @faznet: Der Panzerhersteller Krauss-Maffei Wegmann besiegelt den Zusammenschluss mit dem französischen Rüstungskonzern Nexter http://t.…
- smartdata/tweet_209.xml: BHF Kleinwort Benson: Übernahme durch Privatbank in trockenen Tüchern...

Future / Possible / Planned Event or Relation

Annotate as a relation mention, including trigger. Rename file to include suffix "_future.xml". Note: Some events may only occur after a "green-lighting"/authorization from some authoritative body, e.g. large mergers. In this case, an event counts as future and not yet factual if it has only passed this authorization stage (e.g. Edeka-Tengelmann Fusion below).

Examples

- RT @rpo_wirtschaft: Eine überraschende Wende: #Apple-Zulieferer #Foxconn zögert bei #Sharp-Übernahme https://t.co/V9uwYqn70K
- RT @SPIEGELONLINE: Größer als BASF: US-amerikanische Chemie-Unternehmen DuPont und Dow Chemical planen Mega-Fusion. https://t.co/NLcTgzuXPD
- smartdata/tweet_212_future.xml: RT @SZ: Bundeswirtschaftsminister Sigmar Gabriel billigt Edeka-Tengelmann-Fusion - mit diesen harten Auflagen https://t.co/7ccQ6lZvJy
- smartdata/tweet_220_future.xml: Freie Bahn für Zusammenschluss der Volkshochschulen https://t.co/03XfdQdri2
- smartdata/tweet_218_future.xml: RT @SPIEGELONLINE: Der Verlag @axelspringer und der TV-Sender @ProSieben verhandeln offenbar über eine Fusion http://t.co/gBvqb5glj1
- smartdata/tweet_215_future.xml: Mit 25 Jahren Verspätung könnte es zum Zusammenschluss hessischer Rundfunk und ÖR in Thüringen kommen.
- smartdata/tweet_510.xml: @DB_Bahn Bei Facebook kursiert das Gerücht, dass nächste Woche wieder gestreikt wird? Stimmt das????

Negated Event

Do NOT annotate as a relation mention, which also means NOT to annotate a trigger. Rename file to have suffix "_negation.xml".
Texts like "folgende Meldung entfällt" also count as a negation (see eg. smartdata/rss_37.xml)

Examples

- smartdata/tweet_219_negation.xml: RT @sebastianjost: #BHF-Bank: Fosun beteuert, keine Fusion mit Hauck & Aufhäuser anzustreben http://t.co/jdmBCF041p via @welt

"Internal" Relations

Examples

- smartdata/tweet_224.xml Googlu e-Ableger will Staus mit Big Data bekämpfen: Die zum Google-Imperium gehörende Firma Sidewalk Labs hat si... https://t.co/sm6J0jBYtL

"Google-Ableger" - is the “child” Entity, which can be resolved via coref to “Sidewalk Labs”. The token "Google" corresponds to the “parent” entity, "Ableger" is a trigger, and the relation is “SpinOff” (whereas the 2nd sentence expresses a parent-child "CompanyRelationship").

“Imprecise” Relations

Mergers of company divisions / specific industry sectors only:

Examples

- smartdata/tweet_817.xml: Siemens und Gamesa geben Zusammenschluss der Windgeschäfte bekannt https://t.co/jRbPfmfU3O

Here, not Siemens and Gamesa are merging, but only their “wind utility” divisions/businesses. Annotate a merger between “Windgeschäfte” and trigger “Zusammenschluss.

Missing arguments - no relation mentions annotated

- smartdata/tweet_908.xml: 16. Juni 1953: Bauarbeiter der Ost-Berliner Stalinallee treten aus Protest gegen die Erhöhung der Arbeitsnormen in den Streik #unglaublich
- smartdata/tweet_907.xlm: Streik hat begonnen, Lehrer sammeln sich https://t.co/xZN2njcXGh
    - Not a “Strike” event since no company is mentioned, which is currently a required argument
- smartdata/tweet_341.xml: **Info: Der HKX1804 verkehrt heute witterungsbedingt ab Hamburg Hbf. Reisende mit Start in Hamburg Altona nutzen bitte die S-Bahn zum Hbf.
    - not a “canceled stop” since 1st sentence mentions only the location-route, but no other argument - these are in the 2nd sentence (start_loc “Altona”). would require cross-sentence RE
- smartdata/tweet_317.xml: U8: Notarzteinsatz. Es kommt derzeit zu Verspätungen. #BVG
    - The 2nd sentence could be a Delay relation. but is missing a route argument. “Notarzteinsatz” = cause, but not sufficiently explicit for “Delay”.
- smartdata/tweet_272.xml: spiegel.de : Vulkanausbruch auf Java: Bali muss Internationalen Flughafen schließen: Der… http://t.co/d32wN6fk4o
    - 2nd sentence does not contain a “CanceledStop(internationaler Flughafen, schließen)” relation since the required location-route argument is missing

Identity Relation

If necessary for a relation mention with an argument that needs to be resolved via coreference (i.e. the argument entity is referenced with eg a pronoun or noun phrase), annotate the Identity relation for this argument. The identity relation has 2 arguments, “source” (exactly 1 filler) and “target” (1 or more fillers). The first “full” mention of an entity in a document should be used for the source argument, all other, co-referring mentions should be target arguments. If annotating an Identity relation, make sure to annotate all (!) references to this entity in the document!

3 Text Genres

News

Twitter

Annotate entities at the start of a tweet referring to the user/channel a tweet was posted by (i.e. the “source”) ONLY IF the channel corresponds to a company/organization or a person:

Examples

- RT @jungewelt: Feindliche Übernahme #Fraport übernimmt ...
- @rpo_wirtschaft: Eine überrasc…
- RT @ronaldglaeser: Bild berichtet, der Grund
- smartdata/tweet_454.xml: POL-LB: Einbruch in Bietigheim-Bissingen, D

I.e. “jungewelt” and “rpo_wirtschaft” are entities of type organization-company, “ronaldglaeser” of type person, and “POL-LB” of type organization.

Ignore “tokenization” rules in hashtags, i.e. annotate mentions even if they occur without a separator in a hashtag

Examples

- smartdata/tweet_462.xml Borussia Mönchengladbach gegen den FC Sevilla im Borussia Park #BMG #Sev #BMGvSEV #BMGSEV @SevillaFC_ENG @Borussia
    - all occurrences of “BMG” and “SEV”/”Sev”, even in the last hashtag
    - but annotate “SevillaFC_ENG” as the extent in the following user mention, not just “SevillaFC”!

In Tweet texts, not Hashtags/Usermentions: Annotate entity mentions even if they are misspelled / missing white space, e.g. “Paris” in “War neulich inParis im Urlaub”.

If text is too short to determine if a relation/event is explicitly expressed, do not annotate (i.e. annotate Tweets without reading any expanded version of the tweet, e.g. by following an external / t.co link)

Examples (no relation)

- ■ #Hamburg: Grusonstraße - Wöhlerstraße zwischen Bredowbrücke und Liebigstraße ist wegen #Bauarbeiten bis zum 23. Oktober …
    - although we have locations and a “cause” argument (Bauarbeiten), the tweet doesn’t state whether the road is closed or just experiencing slow traffic, ...

Hashtags / Users (Channels)

Mention extent: Annotate without the leading ‘#’ or ‘@’

RSS