Manual annotation complexity grid

Karën Fort

karen.fort@univ-lorraine.fr / https://members.loria.fr/KFort

Institut des sciences du Digital Management & Cognition
COMPOSANTE DE L'UNIVERSITÉ DE LORRAINE

UNIVERSITÉ DE LORRAINE

Dealing with the complexity of manual annotation

Analysing the complexity of an annotation campaign

What to annotate?

How to annotate?

Weight of the context

Synthesis: using the right tool, at the right place

WYMR: What You Must Remember

Manual annotation

Why is it easy/difficult? What part should be helped/automatized?

Parts of speech [Marcus et al., 1993] :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

Manual annotation

Why is it easy/difficult? What part should be helped/automatized?

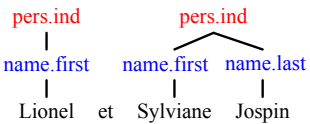
Gene renaming [Fort et al., 2012] :

> *The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU:cat) and recS and "recS1" (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.*

Manual annotation

Why is it easy/difficult? What part should be helped/automatized?

Structured named entities [Grouin et al., 2011] :



What is complex?

Dealing with the complexity of manual annotation

Analysing the complexity of an annotation campaign

What to annotate?

How to annotate?

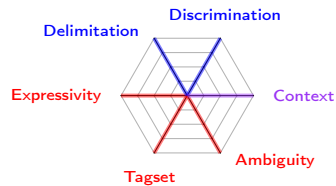
Weight of the context

Synthesis: using the right tool, at the right place

WYMR: What You Must Remember

Complexity dimensions

- 5 independent dimensions:
  - 2 related to the localisation of annotations
  - 3 related to the characterisation of annotations
- 1 not independent: the context



- Scale from **0** (null complexity) to **1** (maximal complexity) to allow for the comparison between campaigns
- Independent from the volume to annotate and the number of annotators

Example: gene renaming

1. Identification of gene names in the source signal:
   *The **yppB** gene complemented the defect of the recG40 strain. **yppB** and **ypbC** and their respective null alleles were termed "**recU**" and "**recU1**" (recU:cat) and "**recS**" and "**recS1**" (recS:cat), respectively.*

2. Identification of gene couples expressing a renaming relation:
   *The **yppB** gene complemented the defect of the recG40 strain. **yppB** and **ypbC** and their respective null alleles were termed "**recU**" and "**recU1**" (recU:cat) and "**recS**" and "**recS1**" (recS:cat), respectively.*

Dealing with the complexity of manual annotation

Analysing the complexity of an annotation campaign

What to annotate?

How to annotate?

Weight of the context

Synthesis: using the right tool, at the right place

WYMR: What You Must Remember

Discrimination

Parts-of-speech [Marcus et al., 1993], pre-annotated :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

Gene renaming[Fort et al., 2012], no pre-annotation:

The yppB:cat and ypbC:cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU:cat) and recS and "recS1" (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.

Discrimination

Parts-of-speech [Marcus et al., 1993], pre-annotated :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

Gene renaming[Fort et al., 2012], no pre-annotation:

The yppB:cat and ypbC:cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU:cat) and recS and "recS1" (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.

$\Rightarrow$ **more difficult** if the units to annotate are scattered, in particular if the segmentation is not obvious.

Discrimination

The discrimination weight is all the more high as the proportion of what *should* be annotated as compared to what *could* be annotated is low.

Definition

$$Discrimination(Flow) = 1 - \frac{|Annotations(Flow)|}{\sum_{i=1}^{LevelSeg} |UnitsObtainedBySeg_i(Flow)|}$$

$\Rightarrow$ Need for a reference segmentation

Parts-of-speech[Marcus et al., 1993] :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

$$Discrimination_{PTB_{POS}} = 0$$

Gene renaming[Fort et al., 2012] :

The yppB:cat and ypbC:cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and "recU1" (recU:cat) and recS and "recS1" (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.

$$Discrimination_{Identification} = 0,9$$
$$Discrimination_{Renaming} = 0,95$$

Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  Madame Chirac $\rightarrow$ *Monsieur et Madame Chirac*

Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  Madame Chirac → *Monsieur et Madame Chirac*
- **decompose** a discriminated unit into several elements:
  le préfet Érignac → *le préfet Érignac*

Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  Madame Chirac → *Monsieur et Madame Chirac*
- **decompose** a discriminated unit into several elements:
  le préfet Érignac → *le préfet Érignac*
- or **group** together several discriminated units into one unique annotation:
  Sa Majesté
  le roi Mohamed VI → *Sa Majesté le roi Mohamed VI*

Boundaries delimitation

Definition

$$Delimitation(Flow) = min\left(\frac{Substitutions + Additions + Deletions}{|Annotations(Flow)|}, 1\right)$$

$$Delimitation_{Identification} = 0$$
$$Delimitation_{Renaming} = 0$$

$$Delimitation_{PTB_{POS}} = 0$$

$$Délimitation_{EN_{TypesSubtypes}} = 1$$
$$Délimitation_{EN_{Components}} = 0,3$$

Dealing with the complexity of manual annotation

Analysing the complexity of an annotation campaign

What to annotate?

How to annotate?

Weight of the context

Synthesis: using the right tool, at the right place

WYMR: What You Must Remember

Expressiveness of the annotation language

Definition

The degrees of expressiveness of the annotation language are the following:

- 0.25: type languages
- 0.5: relational languages of arity 2
- 0.75: relational languages of arity higher than 2
- 1: higher-order languages

$$Expressiveness_{Identification} = 0.25$$
$$Expressiveness_{Renaming} = 0.25$$

Dimension of the tagset

<table>
<tr><th colspan="3">Person</th><th colspan="3">Function</th></tr>
<tr><td colspan="2">pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td colspan="2">func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><th colspan="3">Location</th><th colspan="3">Production</th></tr>
<tr><td rowspan="4">administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td rowspan="4">physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td rowspan="4">facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufactured object)</td><td>prod.serv (transportation route)</td><td>prod.fin (financial products)</td></tr>
<tr><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td colspan="2">prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><td colspan="3"></td><td colspan="3"></td></tr>
<tr><th colspan="3">Organization</th><th colspan="3">Time</th></tr>
<tr><td colspan="2">org.adm (administration)</td><td>org.ent (services)</td><td colspan="2" rowspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td rowspan="2">time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
<tr><th colspan="3">Amount</th></tr>
<tr><td colspan="3">amount (with unit or general object), including duration</td><td colspan="2"></td><td></td></tr>
</table>

Types and sub-types used for structured NE annotation [Grouin et al., 2011]

Dimension of the tagset

<table>
<tr><td colspan="3">Person</td><td colspan="3">Function</td></tr>
<tr><td colspan="2">pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td colspan="2">func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><td colspan="3">Location</td><td colspan="3">Production</td></tr>
<tr><td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td>facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufactured object)</td><td>prod.serv (transportation route)</td><td>prod.fin (financial products)</td></tr>
<tr><td colspan="3"></td><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td colspan="3"></td><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><td colspan="3">Organization</td><td colspan="3">Time</td></tr>
<tr><td colspan="2">org.adm (administration)</td><td>org.ent (services)</td><td colspan="2" rowspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td rowspan="2">time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
<tr><td colspan="3">Amount</td></tr>
<tr><td colspan="3">amount (with unit or general object), including duration</td><td colspan="2"></td><td></td></tr>
</table>

Level 1: *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilities (degree of freedom = 6).

Dimension of the tagset

<table>
<tr><th colspan="3">Person</th><th colspan="3">Function</th></tr>
<tr><td colspan="2">pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td colspan="2">func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><th colspan="3">Location</th><th colspan="3">Production</th></tr>
<tr><td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td>facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufactured object)</td><td>prod.serv (transportation route)</td><td>prod.fin (financial products)</td></tr>
<tr><td colspan="3"></td><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td colspan="3"></td><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><th colspan="3">Organization</th><th colspan="3">Time</th></tr>
<tr><td colspan="2">org.adm (administration)</td><td>org.ent (services)</td><td colspan="2" rowspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td rowspan="2">time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
<tr><th colspan="3">Amount</th></tr>
<tr><td colspan="3">amount (with unit or general object), including duration</td></tr>
</table>

Level 1: *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilities (degree of freedom = 6).

Level 2: *prod.object*, *prod.serv*, *prod.fin*, *prod.soft*, *prod.doctr*, *prod.rule*, *prod.art*, *prod.media*, *prod.award* $\rightarrow$ 9 possibilities (degree of freedom = 8).

Dimension of the tagset

<table>
<tr><th colspan="3">Person</th><th colspan="3">Function</th></tr>
<tr><td colspan="2">pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td colspan="2">func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><th colspan="3">Location</th><th colspan="3">Production</th></tr>
<tr><td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td>facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufactured object)</td><td>prod.serv (transportation route)</td><td>prod.fin (financial products)</td></tr>
<tr><td colspan="3"></td><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td colspan="3"></td><td colspan="2">prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><th colspan="3">Organization</th><th colspan="3">Time</th></tr>
<tr><td colspan="2">org.adm (administration)</td><td>org.ent (services)</td><td colspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td>time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
<tr><th colspan="3">Amount</th><td colspan="3" rowspan="2"></td></tr>
<tr><td colspan="3">amount (with unit or general object), including duration</td></tr>
</table>

Level 1: *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilities (degree of freedom = 6).

Level 2: *prod.object*, *prod.serv*, *prod.fin*, *prod.soft*, *prod.doctr*, *prod.rule*, *prod.art*, *prod.media*, *prod.award* $\rightarrow$ 9 possibilities (degree of freedom = 8).

Level 3: *loc.adm.town*, *loc.adm.reg*, *loc.adm.nat*, *loc.adm.sup* $\rightarrow$ 4 possibilities (degree of freedom = 3).

Dimension of the tagset

Degree of freedom

$$\nu = \nu_1 + \nu_2 + \ldots + \nu_m$$

where ν_i is the maximal degree of freedom the annotator has when choosing the i^{th} sub-type ($\nu_i = n_i - 1$).

Dimension (size) of the tagset

$$Dimension(Flow) = min(\frac{\nu}{\tau}, 1)$$

where τ is the threshold from which we consider the tagset to be very large (experimentally determined).

$$Dimension_{Identification} = 0$$
$$Dimension_{Renaming} = 0.04$$
$$Dimension_{NE_{TypesSubtypes}} = 0.34$$

Degree of ambiguity: residual ambiguity

Using the traces left by the annotators:

[...] <EukVirus>3CDproM</EukVirus> can process both structural and nonstructural precursors of the <EukVirus ***uncertainty-type*** *= "too-generic"><taxon>poliovirus</taxon> polyprotein</EukVirus> [...].*

Définition

$$AmbiguityRes(Flow) = \frac{|Annotations_{amb}|}{|Annotations|}$$

$$AmbiguityRes_{Identification} = 0.04$$
$$AmbiguityRes_{Renaming} = 0.02$$

Degree of ambiguity: theoretical ambiguity

Proportion of the units to annotate that corresponds to ambiguous vocables.

Definition

$$AmbiguityTh(Flow) = \frac{\sum_{voc_i=1}^{|Voc(Flow)|} (Ambig(voc_i) * freq(voc_i, Flow))}{|Units(Flow)|}$$

with

$$Ambig(voc_i) = \begin{cases} 1 & \text{if} \quad |Tags(voc_i)| > 1 \\ 0 & \text{else} \end{cases}$$

$$AmbiguityTh_{Identification} = 0.01$$

$\rightarrow$ Does not apply to renaming relations

Dealing with the complexity of manual annotation

Analysing the complexity of an annotation campaign

What to annotate?

How to annotate?

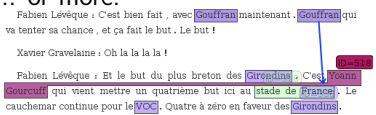
Weight of the context

Synthesis: using the right tool, at the right place

WYMR: What You Must Remember

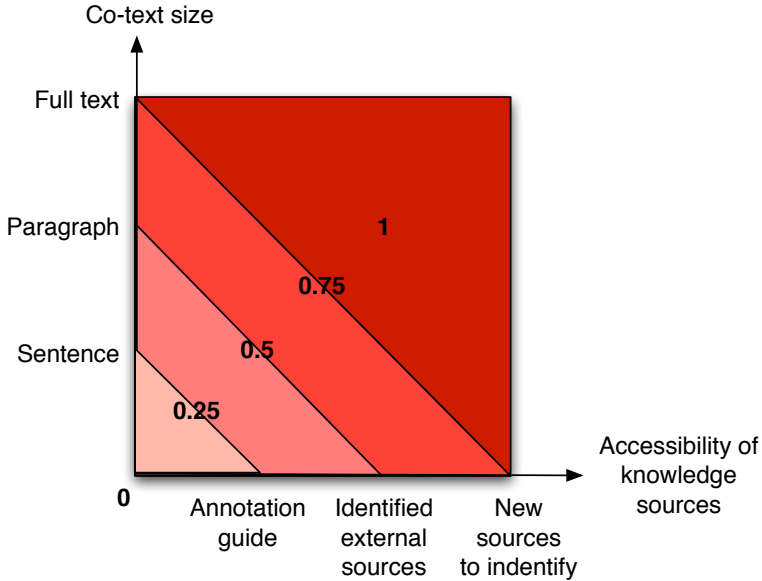
Context to take into account

- **size of the window** to take into account in the source signal:
  - The sentence:
    I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.
  - ... or more:



- number of **knowledge elements** to be rallied or degree of accessibility of the knowledge sources that are consulted:
  - annotation guidelines
  - nomenclatures (Swiss-Prot)
  - new sources to be found (Wikipedia, etc.)

Weight of the context



Dealing with the complexity of manual annotation

Analysing the complexity of an annotation campaign

What to annotate?

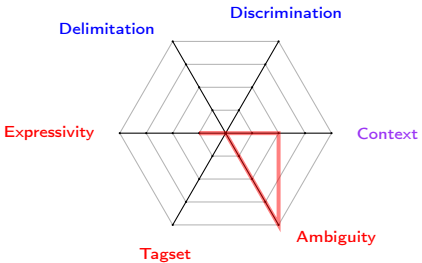
How to annotate?

Weight of the context

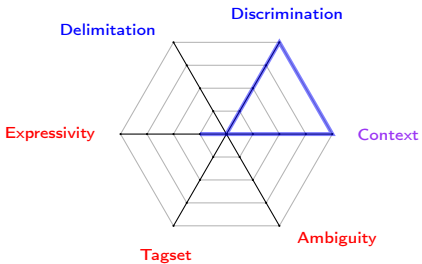
Synthesis: using the right tool, at the right place

WYMR: What You Must Remember

Synthesis of the complexity dimensions



Classification of it pronouns as anaphoric or impersonal



Gene names identification

Manual annotation and NLP:

- usage
- cost

Manual annotation:

- definition
- organization
- complexity grid

Practice: Compute the complexity dimensions for your annotations

Assignment: Reading a research paper (graded)

Read carefully and email your teacher the synopsis (2p max, pdf) by next week

[Dandapat et al., 2009] :
Cheap and Fast – But is it Good?
Evaluating Non-Expert Annotations for Natural Language Tasks
https://aclanthology.org/D08-1027.pdf

Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009).
Complex linguistic annotation - no easy way out! a case from bangla and hindi POS labeling tasks.
In Proceedings of the third ACL Linguistic Annotation Workshop, Singapour.

Fort, K., François, C., Galibert, O., and Ghribi, M. (2012).
Analyzing the impact of prevalence on the evaluation of a manual annotation campaign.
In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turquie.
7 pages.

Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011).
Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview.
In Proceedings of the 5th Linguistic Annotation Workshop, pages 92–100, Portland, Oregon, USA.

Poster.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993).
Building a large annotated corpus of English : The Penn Treebank.
Computational Linguistics, 19(2):313–330.