Recherche Zen
Session 3 : Experiments

Carlos Ramisch and Manon Scholivet

Partly based on the course by Adeline Paiement

18 octobre 2023

Plan

Research question → Experiment

Data annotation

Data quality metrics (agreement)

Data science experiments

Evaluation metrics

Why do we need experiments ?

- A research question and its sub-questions

  → Precise, concise, feasible, interesting
- Hypotheses related to each sub-question
- They are anchored in the litterature and justified

Why do we need experiments ?

- A research question and its sub-questions

  → Precise, concise, feasible, interesting
- Hypotheses related to each sub-question
- They are anchored in the litterature and justified

Experiment goals

1. To build further evidence that will eventually lead to accepting or rejecting the hypothesis
2. Lead to new interesting research questions

Designing an experiment

1. Identify the target hypothesis

   → Prioritise hypotheses according to impact and constraints

Designing an experiment

1. Identify the target hypothesis

   → Prioritise hypotheses according to impact and constraints

2. Identify the needs of the experiment

   → Data, datasets, evaluation metrics

Designing an experiment

1. Identify the target hypothesis

   → Prioritise hypotheses according to impact and constraints

2. Identify the needs of the experiment

   → Data, datasets, evaluation metrics

3. Instantiate under-specified aspects of the question/hypotheses

   → The devil is in the details

Designing an experiment

1. Identify the target hypothesis

    → Prioritise hypotheses according to impact and constraints

2. Identify the needs of the experiment

    → Data, datasets, evaluation metrics

3. Instantiate under-specified aspects of the question/hypotheses

    → The devil is in the details

4. If the result is X, I will be able to conclude Y

    → Reformulate hypotheses in terms of experiment outcomes

Refining the hypothesis : example

Hypothesis

It is possible to learn a model for language L (with no annotations available) from a set of languages L' (with available annotations)

Refining the hypothesis : example

Hypothesis

It is possible to learn a model for language L (with no annotations available) from a set of languages L' (with available annotations)

- A model for which task ? Question answering ? Parsing ?

  → A supervised or unsupervised model ?

- What exact set of languages ?
- What configurations will be tested ?

  → L' contains 1 language, 5 languages...

  → L is similar to a language in L' or not ?

- How to assess if a model is "good" ? Which evaluation metrics ?

Scope

- Experiments in computer science
- Experiments using data
- $\implies$ Experiments in data science

Scope

- Experiments in computer science
- Experiments using data
- $\Longrightarrow$ Experiments in data science

Data science

Is data science a science ?

Scope

- Experiments in computer science
- Experiments using data
- $\Longrightarrow$ Experiments in data science

Data science

Is data science a science ?

Disclaimer : This is not a machine learning course

Experimental protocol

- Step-by-step description of the experiment
- “Algorithm” of the experiment

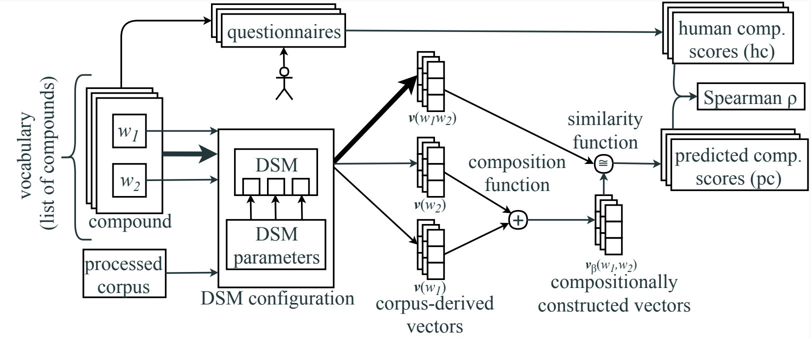
  → Writing the recipe before start cooking

Experimental protocol

How formal is your protocol ?

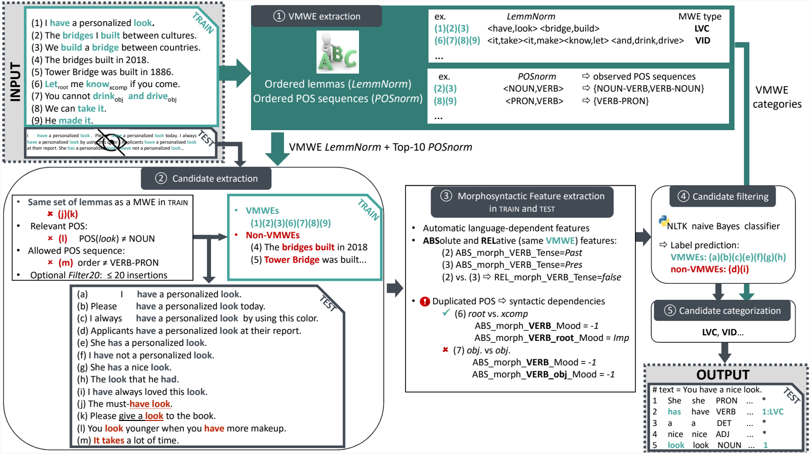
- Depends on the discipline
- A good protocol description can speed up paper writing
- In any case, to be defined before launching experiments

Diagrams to describe the protocol : example 1



Source: `https://aclanthology.org/J19-1001/`

Diagrams to describe the protocol : example 2



INPUT

TRAIN

(1) I **have** a personalized **look**.
(2) The **bridges** I **built** between cultures.
(3) We **build** a **bridge** between countries.
(4) The bridges built in 2018.
(5) Tower Bridge was built in 1886.
(6) **Let**$_{root}$ me **know**$_{xcomp}$ if you come.
(7) You cannot **drink**$_{obj}$ **and drive**$_{obj}$
(8) We can **take it**.
(9) He **made it**.

TEST

I have a personalized look . Please have a personalized look today. I always have a personalized look by using this color. Applicants have a personalized look at their report. She has a personalized look. I have not a personalized look...

① VMWE extraction

Ordered lemmas (*LemmNorm*)
Ordered POS sequences (*POSnorm*)

ex.	*LemmNorm*	MWE type
(1)(2)(3)	<have,look> <bridge,build>	**LVC**
(6)(7)(8)(9)	<it,take><it,make><know,let> <and,drink,drive>	**VID**
...		

ex.	*POSnorm*	⇨ observed POS sequences
(2)(3)	<NOUN,VERB>	⇨ {NOUN-VERB,VERB-NOUN}
(8)(9)	<PRON,VERB>	⇨ {VERB-PRON}
...		

VMWE categories

VMWE *LemmNorm* + Top-10 *POSnorm*

② Candidate extraction

- **Same set of lemmas** as a MWE in TRAIN
  - ✘ **(j)(k)**
- Relevant POS:
  - ✘ **(l)** POS(*look*) ≠ NOUN
- Allowed POS sequence:
  - ✘ **(m)** order ≠ VERB-PRON
- Optional *Filter20*: ≤ 20 insertions

TRAIN

- **VMWEs**
  (1)(2)(3)(6)(7)(8)(9)
- **Non-VMWEs**
  (4) The **bridges built** in 2018
  (5) **Tower Bridge** was built...

TEST

(a) I **have** a personalized **look**.
(b) Please **have** a personalized **look** today.
(c) I always **have** a personalized **look** by using this color.
(d) Applicants **have** a personalized **look** at their report.
(e) She **has** a personalized **look**.
(f) I **have** not a personalized **look**.
(g) She **has** a nice **look**.
(h) The **look** that he **had**.
(i) I **have** always loved this **look**.
(j) The must-**have look**.
(k) Please give a **look** to the book.
(l) You **look** younger when you **have** more makeup.
(m) **It takes** a lot of time.

③ Morphosyntactic Feature extraction in TRAIN and TEST

- Automatic language-dependent features
- **ABS**olute and **REL**ative (same VMWE) features:
  (2) ABS_morph_VERB_Tense=*Past*
  (3) ABS_morph_VERB_Tense=*Pres*
  (2) vs. (3) ⇨ REL_morph_VERB_Tense=*false*
- ! Duplicated POS ⇨ syntactic dependencies
  ✓ (6) *root* vs. *xcomp*
  ABS_morph_**VERB**_Mood = *-1*
  ABS_morph_**VERB_root**_Mood = *Imp*
  ✘ (7) *obj*. vs *obj*.
  ABS_morph_**VERB**_Mood = *-1*
  ABS_morph_**VERB_obj**_Mood = *-1*

④ Candidate filtering

NLTK naive Bayes classifier

⇨ Label prediction:
VMWEs: (a)(b)(c)(e)(f)(g)(h)
non-VMWEs: (d)(i)

⑤ Candidate categorization

LVC, VID...

OUTPUT

TEST

```
# text = You have a nice look.
1   She    she    PRON   ...   *
2   has    have   VERB   ...   1:LVC
3   a      a      DET    ...   *
4   nice   nice   ADJ    ...   *
5   look   look   NOUN   ...   1
```

Source: https://aclanthology.org/W18-4932/

Making choices

- Beware of the combinatorial explosion
  - → # datasets × # configs × # models × # metrics
  - → Grid search = experiments run forever
- Choices must be justified
  - → An arbitrary justification is better than none
  - → E.g. *the parameter was chosen after trial and error*

Making choices

- Favour more promising aspects

  → E.g. Metrics are more or less equivalent $\implies$ choose one

  Datasets are heterogeneous $\implies$ test all of them

  → Small pilot experiments $\implies$ trends $\implies$ choices

Plan

Research question → Experiment

Data annotation

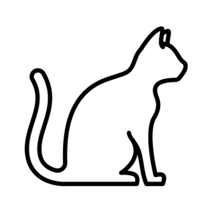
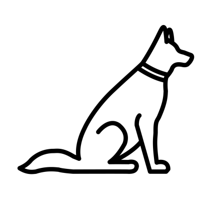
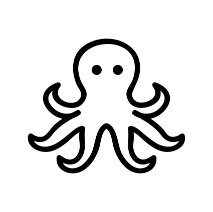
Data quality metrics (agreement)

Data science experiments

Evaluation metrics

Where does data come from ?

- Supervised methods require :
  - Input x + associated gold prediction y

Input			
Reference	Chat	Chien	Poulpe

- gold = reference = label = ground truth

Where does data come from ?

- Machine learning / NLP courses :

  ```
  digits = load_digits()
  print(digits.target[:20]) # magic !
  ```

  ```
  [0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9]
  ```

- Real life :
  - Here's some data (x), apply some learning on it !

    → How to obtain gold/reference labels y to learn/evaluate models ?

Data annotation recipe

1. Select or collect material to annotate
   - licence, biases, representativity, diversity
2. Write annotation guidelines
   - domain expertise, pilot annotation
3. Develop or adapt an annotation platform
   - adaptable, easy to use
4. Train annotators
   - hard cases, speed, biases
5. Evaluate quality
   - inter-rater agreement
6. Combine annotations
   - adjudication, averaging
7. Export and release
   - stable website, format, documentation, articles

Data selection for annotation

- Similarity with target application data
- Trade-off between realistic vs. artificial

  → E.g. newspaper vs. tweets

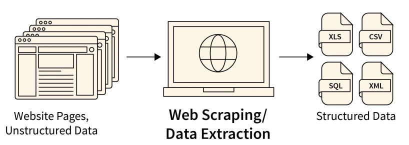
  → Climate crisis means quarter of European ski resorts face scarce snow

  → sooo sick of the snow ughh
- Raw data is noisy $\implies$ harder to annotate/exploit

  → E.g. dialects, typos, code switching, slang

Example : Text crawling / scraping

- Obtain data (HTML) from the web
  - → Off-the-shelf tools, e.g. BootCat
  - → Pre-downloaded web dumps : CommonCrawl, Wikimedia
  - → In-house scripts : parallelisation, `robots.txt`, priority queue, loops



Source: `https://www.scaler.com/topics/data-science/web-scraping/`

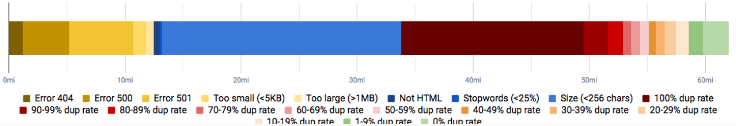
Example : Text crawling / scraping

- Pre-processing and cleaning
  1. Language identification `https://pypi.org/project/langid/`
  2. Deduplication `https://corpus.tools/wiki/Onion`
  3. Text extraction and boilerplate removal

     `https://www.crummy.com/software/BeautifulSoup/`

     `https://pypi.org/project/jusText/`
  4. Content filtering : length, stopword ratio, dictionary
  5. Sentence/word segmentation `https://spacy.io/`

     `https://www.nltk.org/`



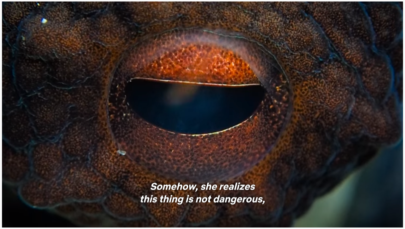
Source: `https://aclanthology.org/L18-1686/`

Indirect annotation

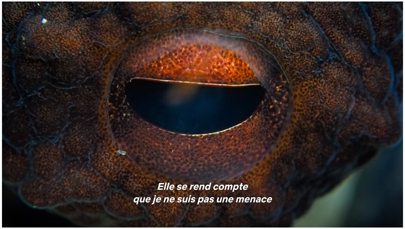
Clever ways to select data

- Open Subtitles : text + translation
    - provided for free by series/movie fans
- Amazon products : text + polarity (positive/negative)
    - Reviews associated with 5-star rating
- Flickr30k : image + description
    - Captions provided by users on Flickr

Example 1 : OpenSubtitles

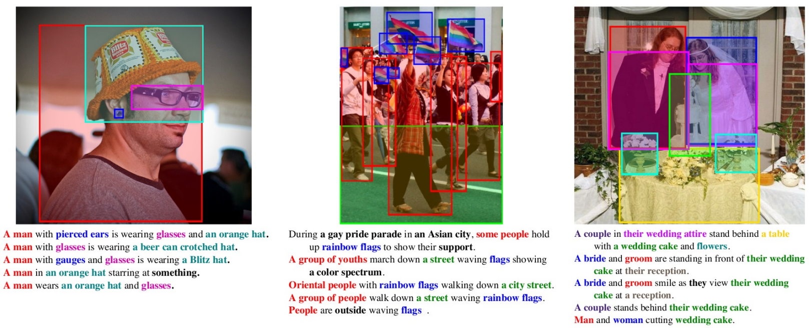


Somehow, she realizes
this thing is not dangerous,



Elle se rend compte
que je ne suis pas une menace

Example 2 : Flickr30k (and extensions)



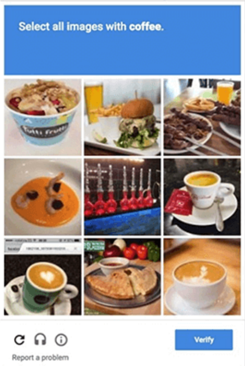
A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .

A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

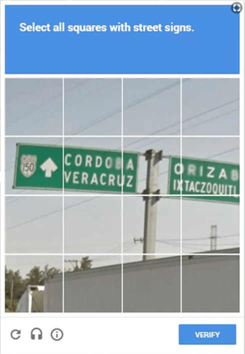
Source: `https://bryanplummer.com/Flickr30kEntities/`

Example 3 : Captcha



Select all images with **coffee**.

First one is a captcha...



Select all squares with street signs.

The second one is free annotation !

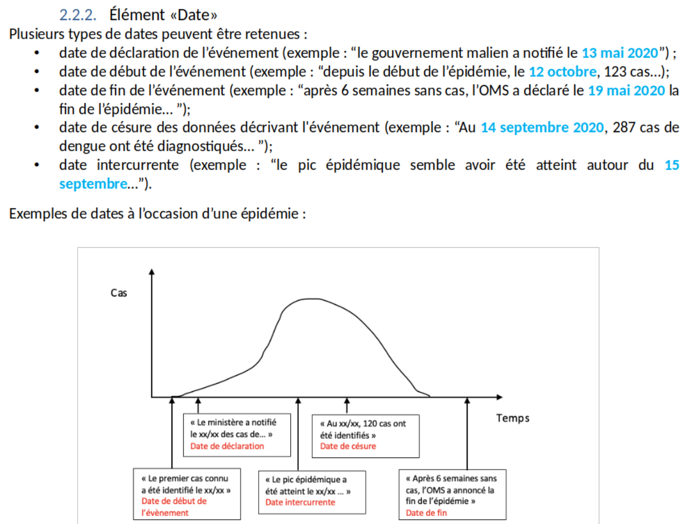
Annotation guidelines

- A document describing the task in much detail
  - → Precise definitions of terms
  - → Homogeneous/standard notation
  - → Describe what may seem obvious
- Describe corner cases
  - → Borderline or difficult phenomena

Annotation guidelines example : epidemiology events

Identify epidemiology events in news

→ date, place, pathology agent, events per document



2.2.2. Élément «Date»

Plusieurs types de dates peuvent être retenues :

- date de déclaration de l'événement (exemple : "le gouvernement malien a notifié le 13 mai 2020") ;
- date de début de l'événement (exemple : "depuis le début de l'épidémie, le 12 octobre, 123 cas…);
- date de fin de l'événement (exemple : "après 6 semaines sans cas, l'OMS a déclaré le 19 mai 2020 la fin de l'épidémie… ");
- date de césure des données décrivant l'événement (exemple : "Au 14 septembre 2020, 287 cas de dengue ont été diagnostiqués… ");
- date intercurrente (exemple : "le pic épidémique semble avoir été atteint autour du 15 septembre…").

Exemples de dates à l'occasion d'une épidémie :

« Le ministère a notifié le xx/xx des cas de… »
Date de déclaration

« Au xx/xx, 120 cas ont été identifiés »
Date de césure

« Le premier cas connu a été identifié le xx/xx »
Date de début de l'évènement

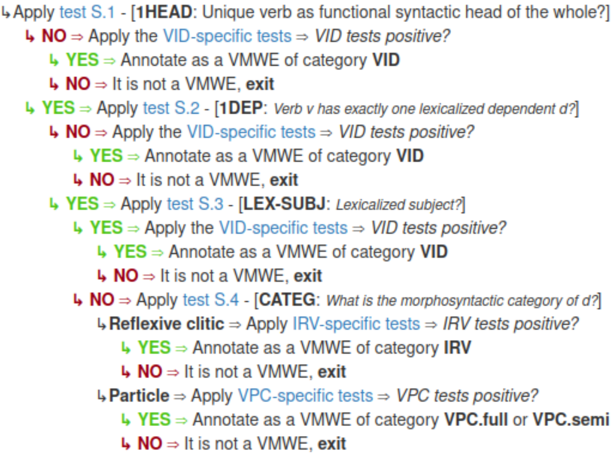
« Le pic épidémique a été atteint le xx/xx … »
Date intercurrente

« Après 6 semaines sans cas, l'OMS a annoncé la fin de l'épidémie »
Date de fin

Annotation guidelines example : multiword expressions

Underline words belonging to multiword expressions

→ span, linguistic criteria, priorities, cross-lingual consistency



- ↳Apply test S.1 - [**1HEAD**: Unique verb as functional syntactic head of the whole?]
  - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
    - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
    - ↳ **NO** ⇒ It is not a VMWE, **exit**
  - ↳ **YES** ⇒ Apply test S.2 - [**1DEP**: *Verb v has exactly one lexicalized dependent d?*]
    - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
      - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
      - ↳ **NO** ⇒ It is not a VMWE, **exit**
    - ↳ **YES** ⇒ Apply test S.3 - [**LEX-SUBJ**: *Lexicalized subject?*]
      - ↳ **YES** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
        - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
        - ↳ **NO** ⇒ It is not a VMWE, **exit**
      - ↳ **NO** ⇒ Apply test S.4 - [**CATEG**: *What is the morphosyntactic category of d?*]
        - ↳**Reflexive clitic** ⇒ Apply IRV-specific tests ⇒ *IRV tests positive?*
          - ↳ **YES** ⇒ Annotate as a VMWE of category **IRV**
          - ↳ **NO** ⇒ It is not a VMWE, **exit**
        - ↳**Particle** ⇒ Apply VPC-specific tests ⇒ *VPC tests positive?*
          - ↳ **YES** ⇒ Annotate as a VMWE of category **VPC.full** or **VPC.semi**
          - ↳ **NO** ⇒ It is not a VMWE, **exit**

Source: https://parsemefr.lis-lab.fr/parseme-st-guidelines/

Annotation guidelines example : compositionality

- Given a word combination

  → ***ivory tower*** → privileged situation

- Proportion of whole's meaning predictable from components ?

  $\rightarrow \text{Comp}(\textit{ivory\_tower}, \textit{ivory}, \textit{tower}) = 10\%$

Annotation guidelines example : compositionality

- Given a word combination

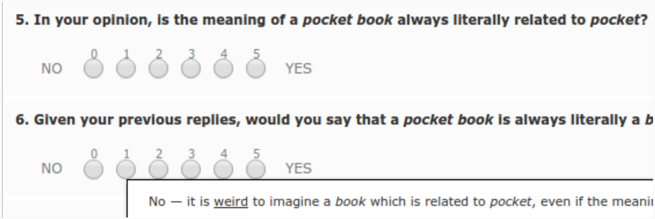
  → ***ivory tower*** → privileged situation

- Proportion of whole's meaning predictable from components ?

  → Comp(*ivory_tower*, *ivory*, *tower*) = 10%

- Scale from 0 (totally idiomatic) to 5 (totally compositional)

  → Head (*book*), modifier (*pocket*), compound (*pocket book*)



5. In your opinion, is the meaning of a *pocket book* always literally related to *pocket*?

NO ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 YES

6. Given your previous replies, would you say that a *pocket book* is always literally a *b*

NO ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 YES

No — it is weird to imagine a *book* which is related to *pocket*, even if the meani

How to write (good) guidelines ?

- Always keep in mind : who are the annotators ?
- Pilot annotation phases

  → Versioning and changelogs
- As objective as possible

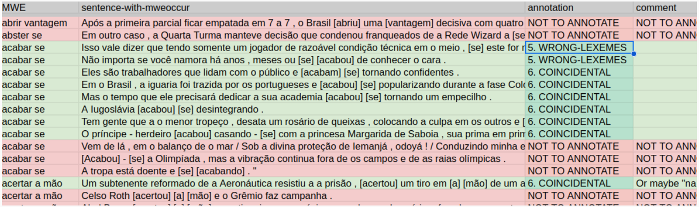
  → Yes/no tests, decision trees, flowcharts
- Cover as many borderline cases as possible

  → Arbitrary but consistent decision, discard if needed
- Add many examples !

  → Explain how to annotate them step by step

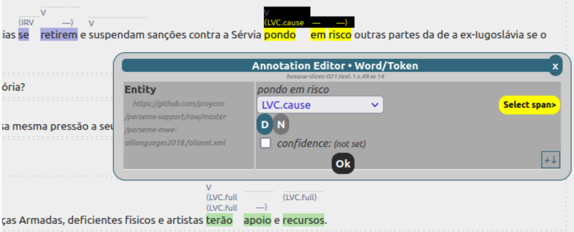
Annotation interface example : DIY

- Generic tools : Excel spreadseets, text files, etc.
- Web forms from scratch : Google forms, PHP, etc.
- Web dev frameworks : Dash, Streamlit, etc.



MWE	sentence-with-mweoccur	annotation	comment
abrir vantagem	Após a primeira parcial ficar empatada em 7 a 7 , o Brasil [abriu] uma [vantagem] decisiva com quatro	NOT TO ANNOTATE	NOT TO ANN
abster se	Em outro caso , a Quarta Turma manteve decisão que condenou franqueados de a Rede Wizard a [se	NOT TO ANNOTATE	NOT TO ANN
acabar se	Isso vale dizer que tendo somente um jogador de razoável condição técnica em o meio , [se] este for r	5. WRONG-LEXEMES	
acabar se	Não importa se você namora há anos , meses ou [se] [acabou] de conhecer o cara .	5. WRONG-LEXEMES	
acabar se	Eles são trabalhadores que lidam com o público e [acabam] [se] tornando confidentes .	6. COINCIDENTAL	
acabar se	Em o Brasil , a iguaria foi trazida por os portugueses e [acabou] [se] popularizando durante a fase Col	6. COINCIDENTAL	
acabar se	Mas o tempo que ele precisará dedicar a sua academia [acabou] [se] tornando um empecilho .	6. COINCIDENTAL	
acabar se	A Iugoslávia [acabou] [se] desintegrando .	6. COINCIDENTAL	
acabar se	Tem gente que a o menor tropeço , desata um rosário de queixas , colocando a culpa em os outros e [	6. COINCIDENTAL	
acabar se	O príncipe - herdeiro [acabou] casando - [se] com a princesa Margarida de Saboia , sua prima em prim	6. COINCIDENTAL	
acabar se	Vem de lá , em o balanço de o mar / Sob a divina proteção de Iemanjá , odoyá ! / Conduzindo minha e	NOT TO ANNOTATE	NOT TO ANN
acabar se	[Acabou] - [se] a Olimpíada , mas a vibração continua fora de os campos e de as raias olímpicas .	NOT TO ANNOTATE	NOT TO ANN
acabar se	A tropa está doente e [se] [acabando] . "	NOT TO ANNOTATE	NOT TO ANN
acertar a mão	Um subtenente reformado de a Aeronáutica resistiu a a prisão , [acertou] um tiro em [a] [mão] de um a	6. COINCIDENTAL	Or maybe "na
acertar a mão	Celso Roth [acertou] [a] [mão] e o Grêmio faz campanha .	NOT TO ANNOTATE	NOT TO ANN

Annotation interface example : FLAT



ias se retirem e suspendam sanções contra a Sérvia pondo em risco outras partes da de a ex-Iugoslávia se o

ória?

sa mesma pressão a seu

ças Armadas, deficientes físicos e artistas terão apoio e recursos.

Annotation Editor • Word/Token

bosque-slices-021.text.1.s.49.w.14

Entity

https://github.com/proycon/parseme-support/raw/master/parseme-mwe-alllanguages2018.foliaset.xml

pondo em risco

LVC.cause

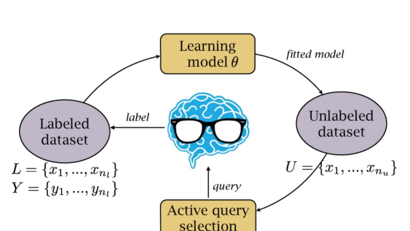
□ confidence: (not set)

Alternatives : Inception, webAnno, brat, FLAT, Arborator, . . .

Automatic pre-annotation

- Pre-annotation
  1. Annotate a small dataset and train predictive model
  2. Predict on the remaining unlabelled data
  3. Correct the predictions

- **Active learning**
  1. Annotate a given instance
  2. Append to training data and train predictive model
  3. Next instance to annotate chosen automatically
     - Maximise diversity of phenomena
     - Maximise the utility for the model

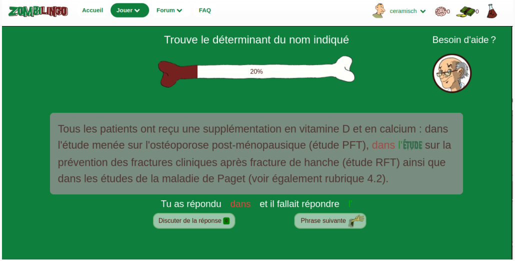


Crowdsourcing

- Compensate for subjectivity = average over many annotators
  - Amazon Mechanical Turk, Crowdflower, ...
- Make the task simpler - accessible for non experts
  - Remuneration per HIT - Human Intelligence Task
- Data quality
  - Qualification pre-task, spammer filtering
- Ethical aspects : unfair remuneration, hard work

Gamification

- Games with a purpose
  - Fun, visually attractive, competition
  - Background : free annotation
- Examples
  - Jeux de mots `https://www.jeuxdemots.org/`
  - ZombiLingo `http://gwap.grew.fr/`



Trouve le déterminant du nom indiqué

Tous les patients ont reçu une supplémentation en vitamine D et en calcium : dans l'étude menée sur l'ostéoporose post-ménopausique (étude PFT), dans l'ÉTUDE sur la prévention des fractures cliniques après fracture de hanche (étude RFT) ainsi que dans les études de la maladie de Paget (voir également rubrique 4.2).

Tu as répondu dans et il fallait répondre l'

Plan

Research question → Experiment

Data annotation

Data quality metrics (agreement)

Data science experiments

Evaluation metrics

Double annotation protocol

- Two (expert/trained) annotators :
  - same training, same annotation guidelines
  - annotate the same data
    - no communication while annotating
- Results should be (almost) identical
  - Inter-annotator agreement
  - Adjudication
- High agreement : guide OK, training OK, data quality OK
- Low agreement : restart until high agreement is reached
- "Low" and "High" → Numerical agreement score

Inter-annotator agreement (IAA) : framework

Items, categories and coders :

- Set of *items* : $\{i | i \in I\}$
- Set of *categories* : $\{k | k \in K\}$
- Set of *coders* (annotators) : $\{c | c \in C\}$

Inter-annotator agreement (IAA)

- Simple case : two raters c_1 and c_2
- Observed agreement : proportion of identically annotated items

$$A_O = \frac{1}{|I|} \sum_{k \in K} \delta(n_{1k}, n_{2k})$$

- n_{ik} = number of coders who assigned item i to category k

Observed agreement : example

Item	Annot1	Annot2
1	Green	Blue
2	Blue	Blue
3	Blue	Green
4	Green	Green
5	Blue	Blue
6	Blue	Blue
	...	...

Contingency table

	Green	Blue	Total
Green	41	3	44
Blue	9	47	56
Total	50	50	100

$$A_O = \frac{1}{|I|} \sum_{k \in K} \delta(n_{1k}, n_{2k})$$

Observed agreement : example

Item	Annot1	Annot2
1	Green	Blue
2	Blue	Blue
3	Blue	Green
4	Green	Green
5	Blue	Blue
6	Blue	Blue
	...	...

Contingency table

	Green	Blue	Total
Green	41	3	44
Blue	9	47	56
Total	50	50	100

$$A_O = \frac{1}{|I|} \sum_{k \in K} \delta(n_{1k}, n_{2k})$$

$$= \frac{41 + 47}{100} = 0.88$$

Adapted from Ron Artstein's slides :

http://ron.artstein.org/publications/2012-artstein-agreement-slides.pdf

Chance-corrected agreement

Task : diagnose whether patients are ill

	Healthy	Ill	Total
Healthy	990	5	995
Ill	5	0	5
Total	995	5	1000

$$A_O = \frac{990}{1000} = 0.99$$

- Most patients are not ill
  - No agreement in ill" category
- High expected agreement A_E
  - How to estimate A_E ?

Cohen's kappa inter-annotator agreement

- Proportion of agreement above chance

$$\kappa = \frac{A_O - A_E}{1 - A_E}$$

- Assume each annotator has their distribution (Cohen's κ)

$$A_E = \frac{1}{|I|^2} \sum_{k \in K} n_{c_1 k} n_{c_2 k}$$

- |I| annotated items in total,
- K possible values per item,
- $n_{c_j k}$ items annotated as k by rater c_j

Adapted from Ron Artstein's slides :

http://ron.artstein.org/publications/2012-artstein-agreement-slides.pdf

Exercise : calculate kappa

	Healthy	Ill	Total
Healthy	990	5	995
Ill	5	0	5
Total	995	5	1000

- $|I| = 1000$ annotated items in total,
- $n_{c_j k}$ items annotated as k by rater c_j

$$A_O = \frac{990}{1000} = 0.99 \qquad \kappa = \frac{A_O - A_E}{1 - A_E} \qquad A_E = \frac{1}{|I|^2} \sum_{k \in K} n_{c_1 k} n_{c_2 k}$$

1. Calculate the kappa chance-corrected IAA score

Exercise : calculate kappa

	Healthy	Ill	Total
Healthy	990	5	995
Ill	5	0	5
Total	995	5	1000

- $|I| = 1000$ annotated items in total,
- $n_{c_j k}$ items annotated as k by rater c_j

$$A_O = \frac{990}{1000} = 0.99 \qquad \kappa = \frac{A_O - A_E}{1 - A_E} \qquad A_E = \frac{1}{|I|^2} \sum_{k \in K} n_{c_1 k} n_{c_2 k}$$

1. Calculate the kappa chance-corrected IAA score

$$A_E = \frac{995^2 + 5^2}{1000^2} = 0.995^2 + 0.005^2 = 0.99005 \qquad A_O = 0.99 \qquad \kappa = -0.005$$

More complex cases

- More than 2 raters
  - Consider pairs of agreeing annotators

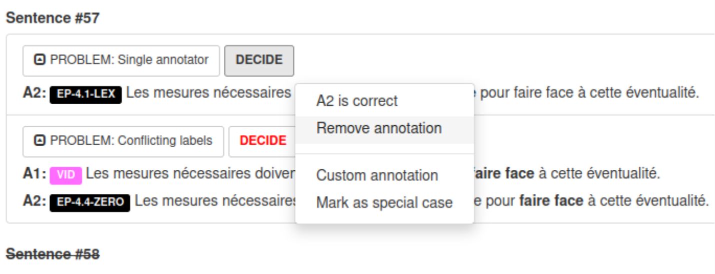
    → Fleiss' kappa

    → Alpha coefficient (take into acccount distance between categories)
- Sporadic annotations
  - F-score between raters

Source: Further reading - `https://aclanthology.org/J08-4004/`

Adjudication

- Carried out by another expert (not an annotator)
- Dedicated interface
- Documented conflict resolution strategies



Sentence #57

A2: EP-4.1-LEX Les mesures nécessaires … pour faire face à cette éventualité.

A1: VID Les mesures nécessaires doiven… **aire face** à cette éventualité.

A2: EP-4.4-ZERO Les mesures nécessaires … e pour **faire face** à cette éventualité.

~~**Sentence #58**~~

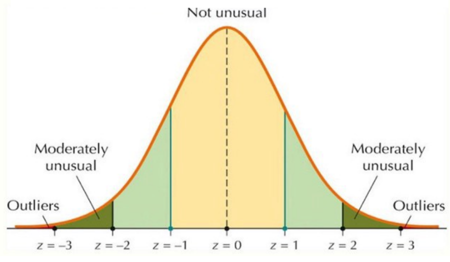
- Creation of final (adjudicated) dataset

Data cleaning

- Some annotations are outliers
- Cleaning must occur before experiments

Z-score filtering

Remove annotations that are more than z standard deviations away from the mean



Source: Further reading : `https://aclanthology.org/W16-1804/`

Data splitting i

- Evaluation must be carried out on held out data

  → Test dataset

- Development must be carried out on held out data

  → Development or validation dataset

  → Attention : it is extremely easy to accidentally tune on test data

- Parameters must be learned from data

  → Training dataset

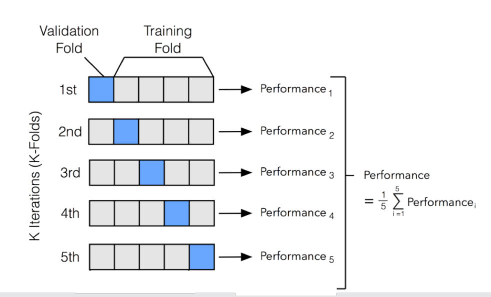
Data splitting ii

Fixed split

- Randomly pick 10% for test, 10% for dev, 80% for train
- Comparable across experiments, papers

Data splitting iii

k-fold cross validation



- Expensive : requires training k models instead of 1

Data splitting iv

Biased split

- Fixed split, but not random
- The test set has controlled characteristics

  → E.g. test instances are unseen in training data

Data splitting v

Discussion

- *We need to talk about standard splits*

  → `https://aclanthology.org/P19-1267/`

- *We need to talk about random splits*

  → `https://aclanthology.org/2021.eacl-main.156/`

- . . .

Understand the data

- Open your files !

  → Otherwise someone may troll you :

  https://medium.com/@yoav.goldberg/

  an-adversarial-review-of-adversarial-generation-of-natural-language-409ac

- Don't try to get blood from a turnip

  → Maybe your prediction task is unrealistic

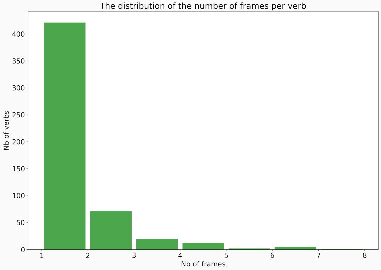
  → Maybe you need external resources

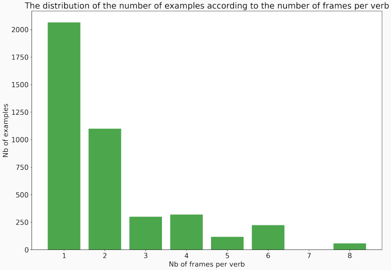
  → ...

Data analysis

- Distribution of classes, input characteristics
- Useful tool : histogram (e.g. `matplotlib.pyplot.hist`)

  → Use bins to discretise real-valued attributes





Source: Author : Anna Mosolova

Benchmarks

- Use benchmarks to compare your method with others

  → Questions about the quality of standard datasets

- Shared tasks :

  → Help make progress, but

  → Encourage using low-quality data for years and years for the sake of comparability

Annotation beyond dataset creation

- Annotating = understanding your problem

  → Hard for humans ? $\Longrightarrow$ maybe hard for models

  → Low agreement $\Longrightarrow$ maybe ill-defined problem

  → Annotation guidelines $\Longrightarrow$ inspiration for features

Plan

Research question → Experiment

Data annotation

Data quality metrics (agreement)

Data science experiments

Evaluation metrics

Experimental conditions

- Supervised, unsupervised, semi-supervised
- Generalisation and amount of supervision

  → Zero-shot, one-shot, few-shot
- Model's (hyper-)parameters

  → E.g. Neural network architecture, dimensions, . . .

  → E.g. Clustering linking criterion, threshold

Baseline and topline i

- A model is never good or bad per se
- Situate the model performance wrt. a simpler model
  - → Baseline – simple model for the task
- Examples of baseline
  - → Random prediction
  - → Majoritary class
  - → A good model 5 years ago
  - → An interpretable model (rules, thresholds)
  - → State-of-the-art model published last month

Baseline and topline ii

- Situate the model performance wrt. a better model

  → Topline – upper bound for the performance

- Examples of topline

  → State-of-the-art model published last month

  → Large model released by big tech company

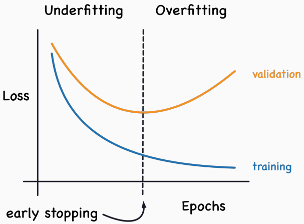
  → Human annotator performance/agreement

  → Same experiment in unrealistic (easy) condition

Overfitting

- The model "overfits" if it memorises the training set
- Tools to prevent overfitting
  - Rule of thumb of pre-neural models :

    → Less features than data items
  - Learning curves on dev set
  - Early stopping based in dev set performance



Hyperparameter search

- Some important hyperparameters
  - learning rate
  - epochs/early stopping patience
  - batch size
  - dropout ratios
  - model capacity (hidden layer dimensions)
  - number of stacked layers, attention heads
  - embedding size
- Tuning strategies
  - Grid search
  - Bayesian search
  - Random search
  - ...
- Unavoidable but usually not very interesting

Model instability

- Same hyperparameters, different random seeds
  - weight initialisaiton in fine-tuning layers
  - order of inputs/batches
- Substantially different results
  - Some data orders/initializations consistently better than others
  - Early stoppin is effective
- **Report averages, error bars, confidence intervals**
  - Re-run training several times with different orders/random initialisation seeds
  - Explicitly set `random.seed` (for each lib), record and publish values

Source: Further reading : `https://arxiv.org/abs/2002.06305`

For later...

- Experiments management
- Reproducibility vs. replicability

Plan

Research question → Experiment

Data annotation

Data quality metrics (agreement)

Data science experiments

Evaluation metrics

Disclaimer : all metrics are incomplete

- Ideally : measure a hidden variable or phenomenon
- In practice : measure what we can observe

  → Formulation is simple enough to be interpretable
- Metrics are partial views of the results

Classification framework

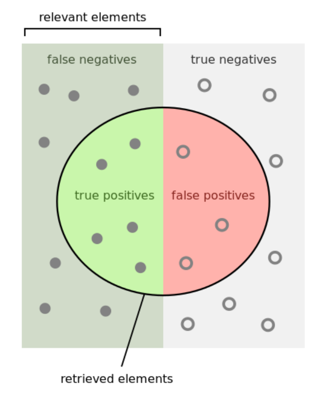
- *tp* : True Positives

  → Correctly predict as positive
- *tn* : True Negatives

  → Correctly predict as negative
- *fp* : False Positives

  → Predict positive, should be negative
- *fn* : False Negatives

  → Predict negative, should be positive



Source: Image : Wikipedia

Accuracy

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$$

- Percentage of well classified items
- Incomplete description of the method's performance

[Image : Devin Soni, towardsdatascience.com]

Precision, recall, F-score

- Calculated per predicted category
- Precision/recall : Complementary measures, report both !
  - Precision

    $\rightarrow tp/(tp + fp)$
  - Recall = Sensitivity

    $\rightarrow tp/(tp + fn)$
  - Specificity :

    $\rightarrow tn/(tn + fp)$
- F-score : Harmonic mean of precision and recall

$$F = 2.\frac{precision.recall}{precision+recall}$$

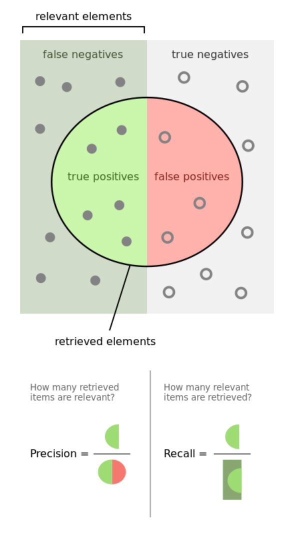


Image from Wikipedia

Accuracy and class imbalance

- Example : hate speech detection in tweets
  - Only a small percentage (∼1%) are hateful
  - Let's annotate everything as not hateful
  - My model has an accuracy of 99% ! So powerful !

F-score or F-measure

- F-score (or F-measure) : harmonic mean of precision and recall

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

- F-score can be weighted to favour precision or recall

  → β=0.5 : More weight on precision, less weight on recall

  → β=1 : Balance the weight on precision and recall

  → β=2 : Less weight on precision, more weight on recall

$$F_\beta = (1 + \beta^2) \times \frac{precision \times recall}{\beta^2 precision + recall}$$

- Does it make sense to average F-scores ?

  → Macro- or micro-average ?

Other metrics (see backup slides)

- ROC curve / Area under the curve

  → Real prediction, threshold

- (Mean) average precision

  → Real prediction, binary gold classes

- Structured prediction

  → Compare trees, graphs

- ...

Goodhart's law

"When a measure becomes a target, it ceases to be a good measure"

- Cobra effect
- Reinforcement learning policies
- Grade-oriented education system

- Risk : optimise evaluation metric at any expense

  → Overfitting, low generalisation

  → Forgetting the research question

  → Frustration with unrealistic goals

  → . . .

Source: Thanks to François Hamonic for this slide.

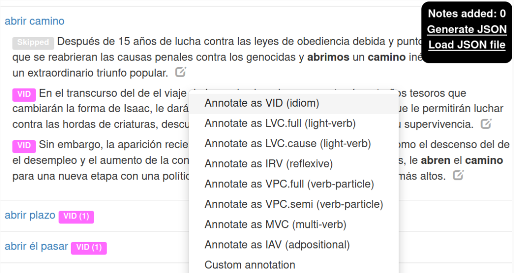
Sources

- Cours d’Adeline Paiement
- Wikipedia
- Google images

Backup slides

Consistency checks

- Vertical data visualisation
  - Aggregate similar units (e.g. by lemma, POS n-gram, etc)
- Adjudicator of expert annotator corrects mistakes



abrir camino

Skipped Después de 15 años de lucha contra las leyes de obediencia debida y punto que se reabrieran las causas penales contra los genocidas y **abrimos** un **camino** iné un extraordinario triunfo popular.

VID En el transcurso del de el viaje os tesoros que cambiarán la forma de Isaac, le dará ue le permitirán luchar contra las hordas de criaturas, descu u supervivencia.

VID Sin embargo, la aparición recie omo el descenso del de el desempleo y el aumento de la con s, le **abren** el **camino** para una nueva etapa con una polític más altos.

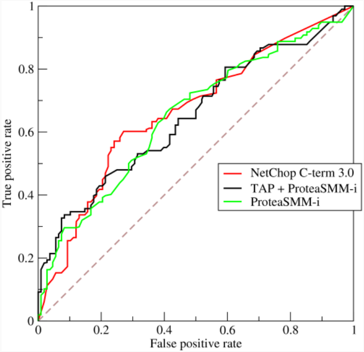
abrir plazo VID (1)

abrir él pasar VID (1)

Notes added: 0
Generate JSON
Load JSON file
Annotate as VID (idiom)
Annotate as LVC.full (light-verb)
Annotate as LVC.cause (light-verb)
Annotate as IRV (reflexive)
Annotate as VPC.full (verb-particle)
Annotate as VPC.semi (verb-particle)
Annotate as MVC (multi-verb)
Annotate as IAV (adpositional)
Custom annotation

ROC curve

ROC curves (*Receiver Operating Characteristic*) are very useful to chose a threshold.

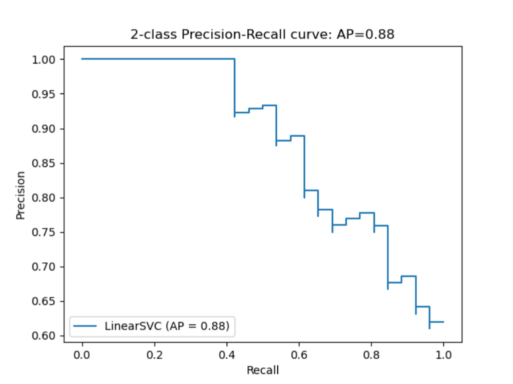


The AUC (*Area Under ROC*) is often used to estimate the model skill.

Image from Wikipedia

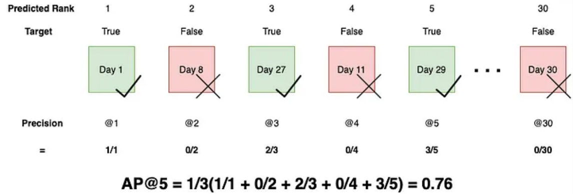
Precision-recall curve

Another way to do this is to use the Precision and the Recall instead of using the True positive and the False positive rates.



Mean average precision

- Model predicts a numerical score
- Gold class is binary or discrete
- Evaluate without setting a fixed threshold



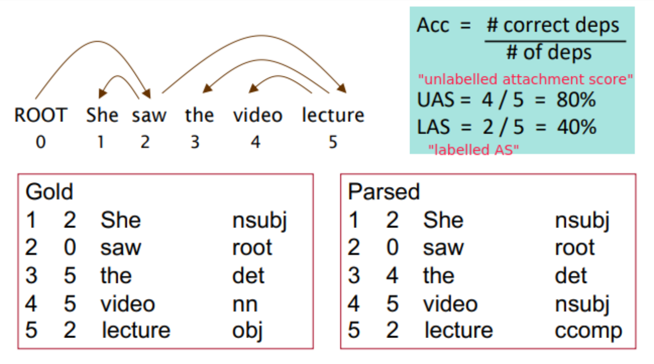
Predicted Rank	1	2	3	4	5		30
Target	True	False	True	False	True		False
	Day 1 ✓	Day 8 ✗	Day 27 ✓	Day 11 ✗	Day 29 ✓	. . .	Day 30 ✗
Precision	@1	@2	@3	@4	@5		@30
=	1/1	0/2	2/3	0/4	3/5		0/30

AP@5 = 1/3(1/1 + 0/2 + 2/3 + 0/4 + 3/5) = 0.76

Structured prediction

- How to compare structured objects ?
  - → Sub-sequences
  - → Clusters
  - → Syntax trees
  - → Graphs

Structured prediction example : LAS/UAS



$$\text{Acc} = \frac{\text{\# correct deps}}{\text{\# of deps}}$$

"unlabelled attachment score"

UAS = 4 / 5 = 80%

LAS = 2 / 5 = 40%

"labelled AS"

Gold

1	2	She	nsubj
2	0	saw	root
3	5	the	det
4	5	video	nn
5	2	lecture	obj

Parsed

1	2	She	nsubj
2	0	saw	root
3	4	the	det
4	5	video	nsubj
5	2	lecture	ccomp

Source: `https://x-wei.github.io/xcs224n-lecture5.html`