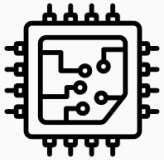
Is NLP Ready for Standardization?

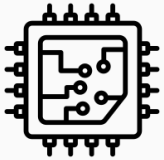
Lauriane Aufrant

Findings of EMNLP 2022

What is the common point between...



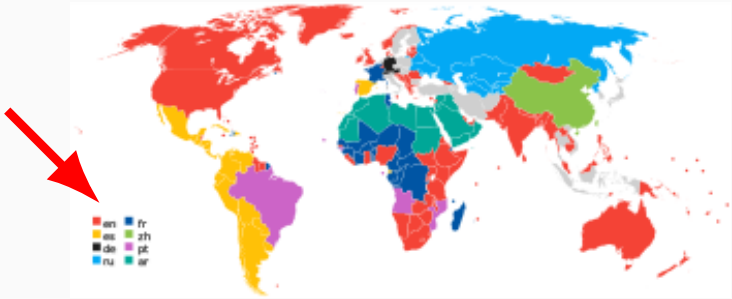
What is the common point between...



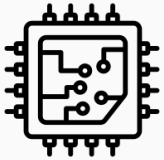
THE C PROGRAMMING LANGUAGE

mp3

```
 !"#$%&'()*+,-./0123456789:;<=>?
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_
`abcdefghijklmnopqrstuvwxyz{|}~
 ¡¢£¤¥¦§¨©ª«¬-®¯°±²³´µ¶·¸¹º»¼½¾¿
ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
àáâãäåæçèéêëìíîïðñòóôõö÷øùúûüýþÿ
```



What is the common point between...



ISO/IEC 9899

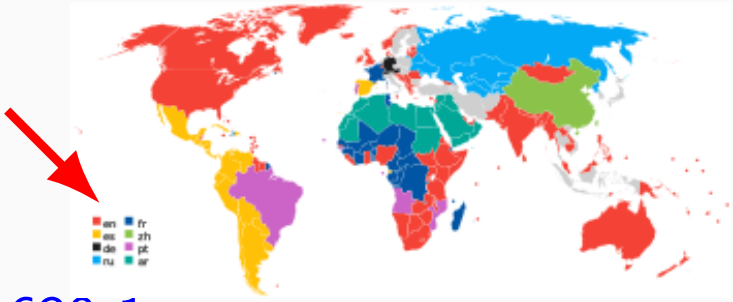
THE C PROGRAMMING LANGUAGE

mp3

ISO/IEC 11172-3

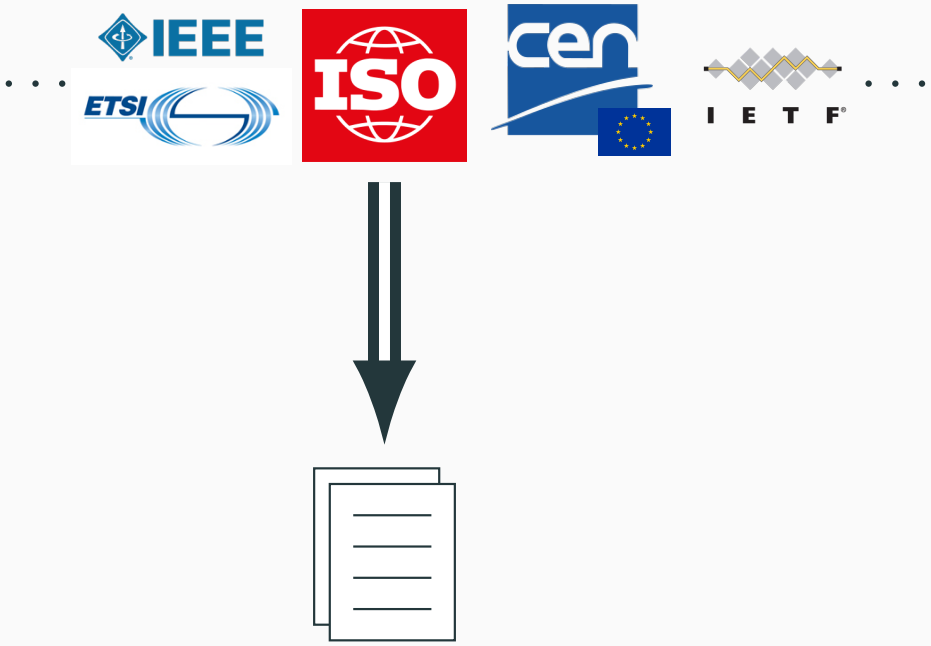
ISO/IEC 8859-1

```
 !"#$%&'()*+,-./0123456789:;<=>?
@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_
`abcdefghijklmnopqrstuvwxyz{|}~
 ¡¢£¤¥¦§¨©ª«¬-®¯°±²³´µ¶·¸¹º»¼½¾¿
ÀÁÂÃÄÅÆÇÈÉÊËÌÍÎÏÐÑÒÓÔÕÖ×ØÙÚÛÜÝÞß
àáâãäåæçèéêëìíîïðñòóôõö÷øùúûüýþÿ
```

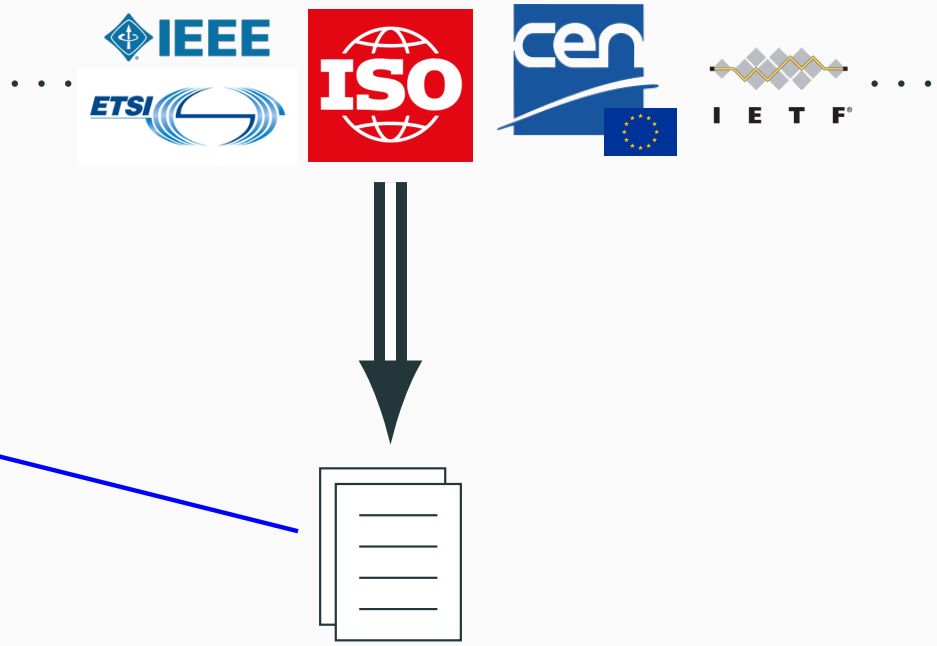


ISO 639-1

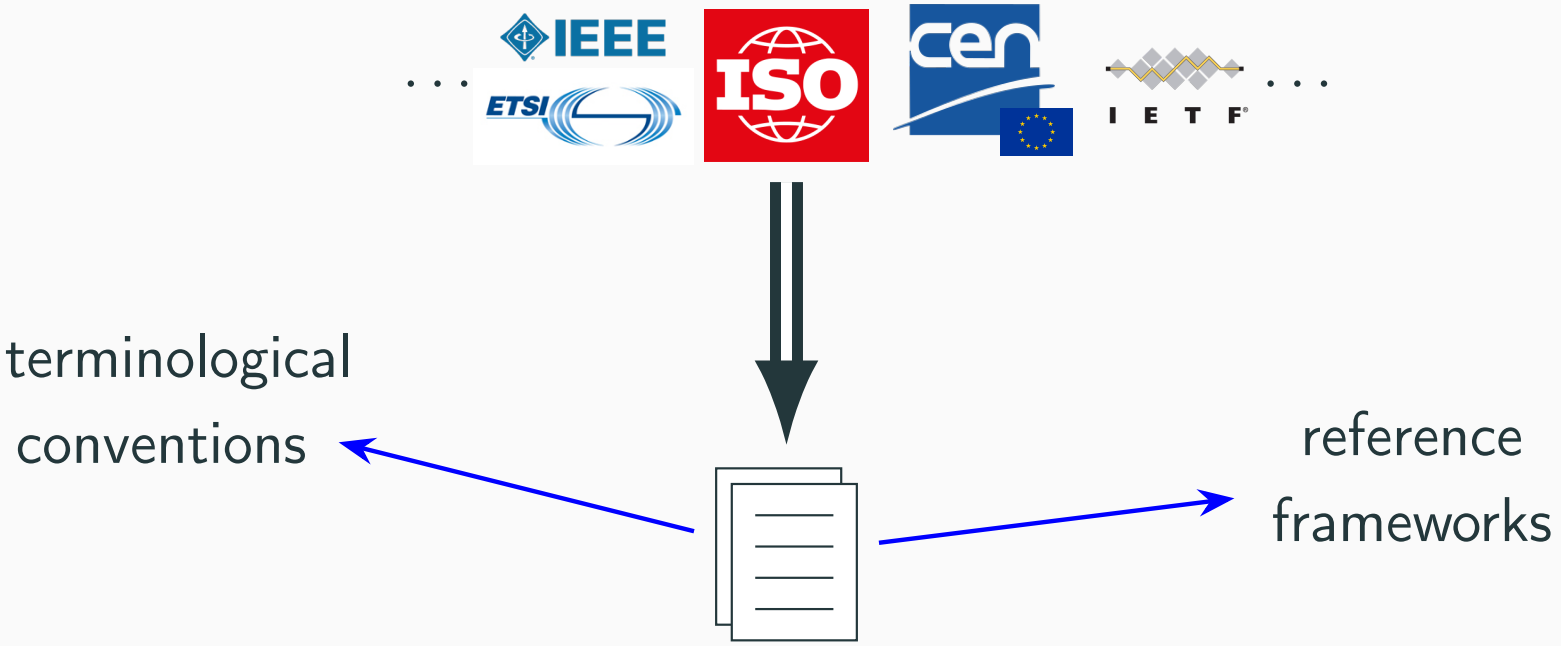
Standards



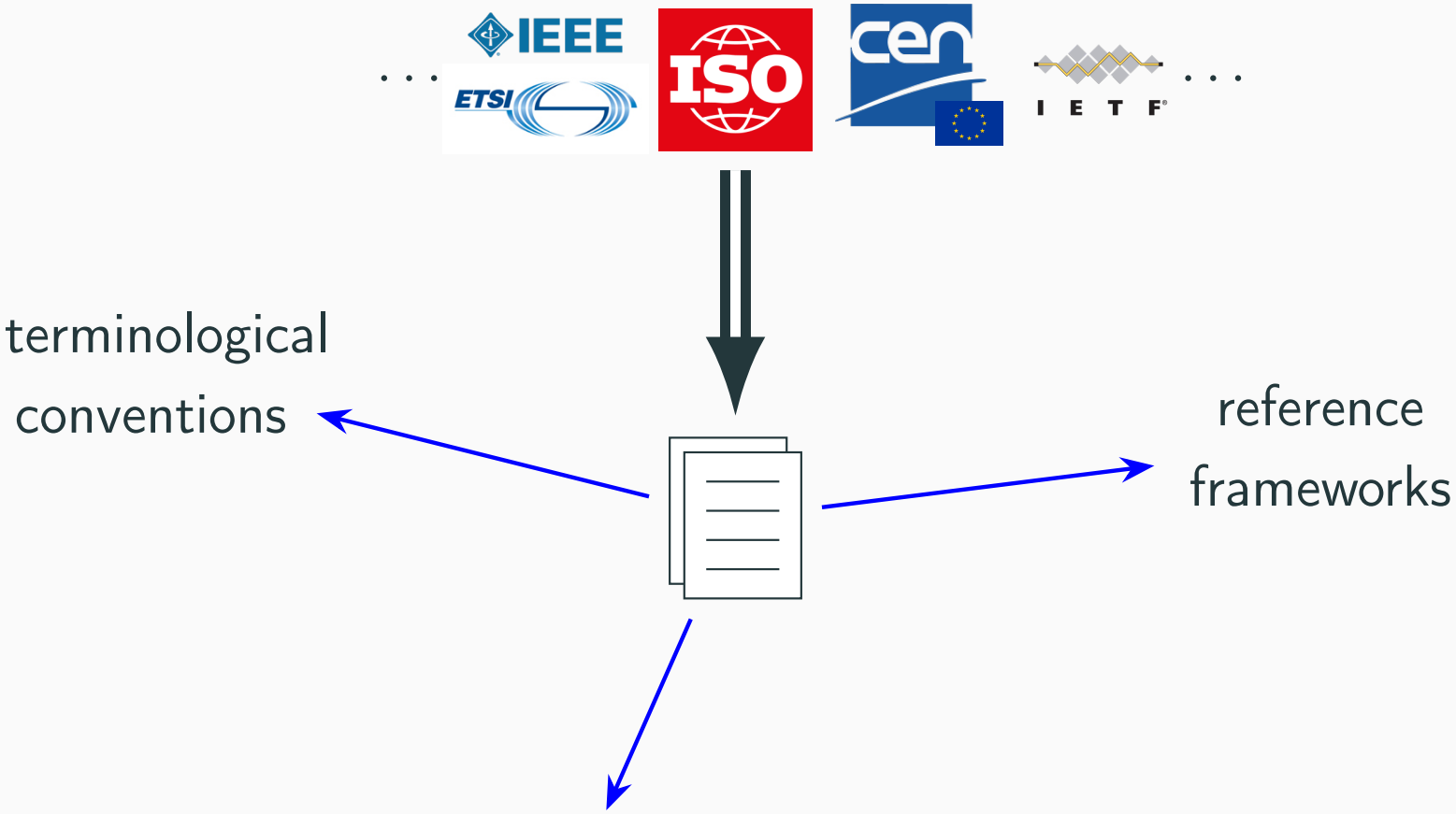
Standards



Standards


IEEE
ETSI
ISO
cen
IETF
terminological
conventions
reference
frameworks

Standards

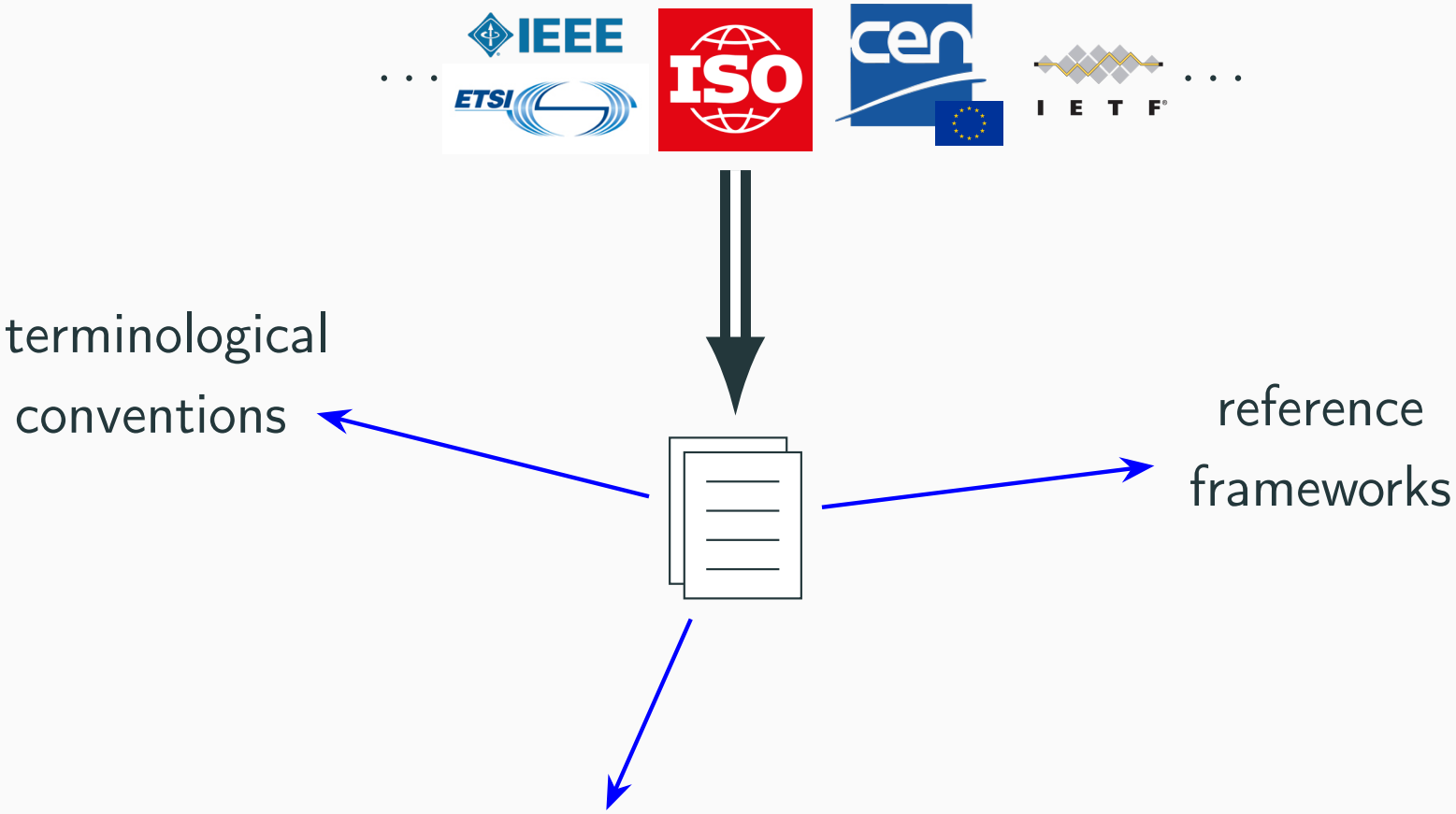


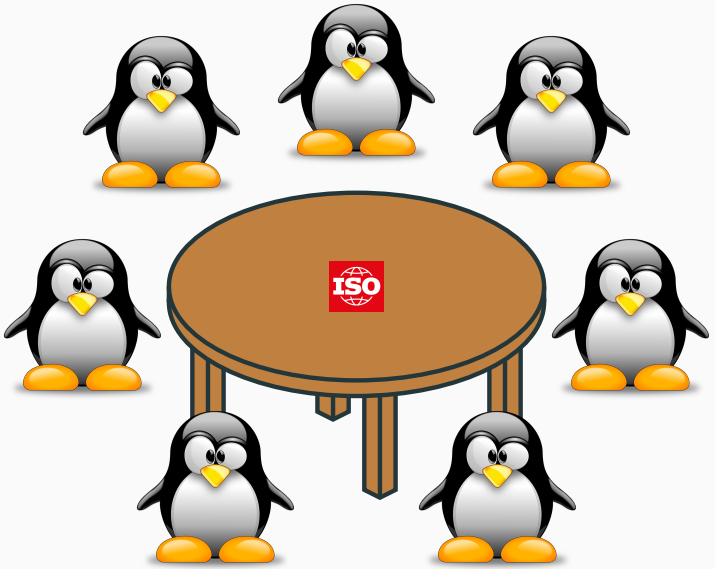
technical specifications

↪ for data, systems, procedures...

↪ quality metrics, interoperability (APIs, protocols...)

Standards


IEEE
ETSI
ISO
cen
IETF
terminological
conventions
reference
frameworks
technical specifications
compliance test suites
certification
↪ for data, systems, procedures...
↪ quality metrics, interoperability (APIs, protocols...)


ISO


ISO


ISO

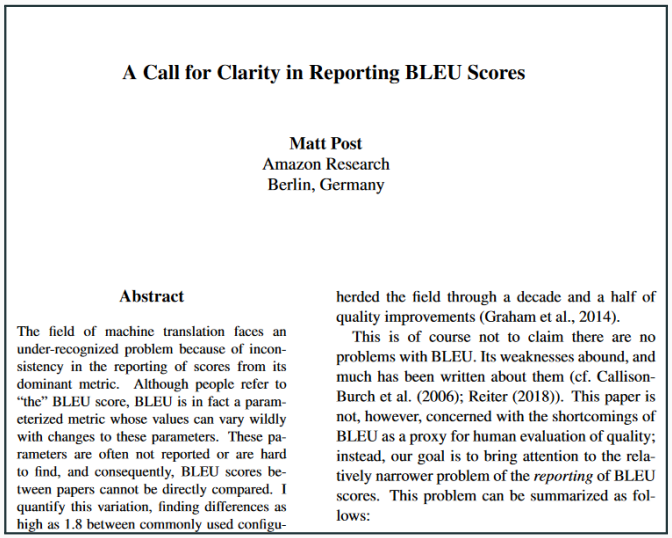
And... NLP?

Towards fair and reproducible evaluation: Machine translation & BLEU

$$\mathrm{BLEU} = \exp^{-\max\left(0, \frac{L}{L_{\textit{closest-ref}}} - 1\right)} \cdot \sqrt[n]{\prod_{k=1}^{n} \frac{\mathrm{TP}_{\textit{n-gram}}(k)}{\mathrm{TP}_{\textit{n-gram}}(k) + \mathrm{FP}_{\textit{n-gram}}(k)}}$$

Towards fair and reproducible evaluation: Machine translation & BLEU

$$\text{BLEU} = \exp^{-\max(0, \frac{L}{L_{\textit{closest-ref}}} - 1)} \cdot \sqrt[n]{\prod_{k=1}^{n} \frac{\text{TP}_{\textit{n-gram}}(k)}{\text{TP}_{\textit{n-gram}}(k) + \text{FP}_{\textit{n-gram}}(k)}}$$



A Call for Clarity in Reporting BLEU Scores

Matt Post
Amazon Research
Berlin, Germany

Abstract

The field of machine translation faces an under-recognized problem because of inconsistency in the reporting of scores from its dominant metric. Although people refer to "the" BLEU score, BLEU is in fact a parameterized metric whose values can vary wildly with changes to these parameters. These parameters are often not reported or are hard to find, and consequently, BLEU scores between papers cannot be directly compared. I quantify this variation, finding differences as high as 1.8 between commonly used configu-

herded the field through a decade and a half of quality improvements (Graham et al., 2014).

This is of course not to claim there are no problems with BLEU. Its weaknesses abound, and much has been written about them (cf. Callison-Burch et al. (2006); Reiter (2018)). This paper is not, however, concerned with the shortcomings of BLEU as a proxy for human evaluation of quality; instead, our goal is to bring attention to the relatively narrower problem of the *reporting* of BLEU scores. This problem can be summarized as follows:

- Tokenization (user's vs metric-internal)?
- N-gram max size?
- Number of references?
- ...

↪ Up to Δ **1.8 BLEU** ($>$ gains from BPE!)

Towards fair and reproducible evaluation: Named entity recognition & F1

$$F1 = \frac{2 \cdot P \cdot R}{P + R},\ P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},\ R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Towards fair and reproducible evaluation: Named entity recognition & F1

$$F1 = \frac{2 \cdot P \cdot R}{P + R},\ P = \frac{\mathrm{TP}_?}{\mathrm{TP}_? + \mathrm{FP}_?},\ R = \frac{\mathrm{TP}_?}{\mathrm{TP}_? + \mathrm{FN}_?}$$

Towards fair and reproducible evaluation: Named entity recognition & F1

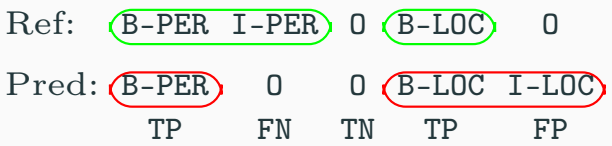
$$F1 = \frac{2 \cdot P \cdot R}{P + R},\ P = \frac{\mathrm{TP}_?}{\mathrm{TP}_? + \mathrm{FP}_?},\ R = \frac{\mathrm{TP}_?}{\mathrm{TP}_? + \mathrm{FN}_?}$$

Ref: B-PER I-PER O B-LOC O

Pred: B-PER O O B-LOC I-LOC

Towards fair and reproducible evaluation: Named entity recognition & F1

$$F1 = \frac{2 \cdot P \cdot R}{P + R}, \; P = \frac{\mathrm{TP}_?}{\mathrm{TP}_? + \mathrm{FP}_?}, \; R = \frac{\mathrm{TP}_?}{\mathrm{TP}_? + \mathrm{FN}_?}$$



↪ 67 F1 on token-level (BIO) labels

↪ 0 F1 on predicted chunks

Towards fair and reproducible evaluation: Named entity recognition & F1

$$F1 = \frac{2 \cdot P \cdot R}{P+R},\ P = \frac{\mathrm{TP}_?}{\mathrm{TP}_? + \mathrm{FP}_?},\ R = \frac{\mathrm{TP}_?}{\mathrm{TP}_? + \mathrm{FN}_?}$$

Ref: `B-PER I-PER` `O` `B-LOC` `O`

Pred: `B-PER` `O` `O` `B-LOC I-LOC`

`TP` `FN` `TN` `TP` `FP`

↪ 67 F1 on token-level (BIO) labels

↪ 0 F1 on predicted chunks

- Typed matching vs boundaries only
- 2 correct boundaries vs 1 correct vs at least 1 correct token
- Non 1-1 matches (ref `B-PER I-PER` vs pred `B-PER B-PER`)

Towards fair and reproducible evaluation: Named entity recognition & F1

$$F1 = \frac{2 \cdot P \cdot R}{P+R},\ P = \frac{\mathrm{TP}_?}{\mathrm{TP}_? + \mathrm{FP}_?},\ R = \frac{\mathrm{TP}_?}{\mathrm{TP}_? + \mathrm{FN}_?}$$

Ref: `B-PER I-PER` `O` `B-LOC` `O`

Pred: `B-PER` `O` `O` `B-LOC I-LOC`

`TP` `FN` `TN` `TP` `FP`

↪ 67 F1 on token-level (BIO) labels

↪ 0 F1 on predicted chunks

- Typed matching vs boundaries only
- 2 correct boundaries vs 1 correct vs at least 1 correct token
- Non 1-1 matches (ref `B-PER I-PER` vs pred `B-PER B-PER`)

And what about...

- [?] invalid sequences `B-PER I-LOC O I-PER O`
- [?] micro vs macro-F1
- [?] ignoring Other/MISC classes
- [?] dependency of evaluation tools to the encoding scheme (BIO, BILUO...)

Consistent data and annotations

	John	**D.**	**Smith**	**arrived**	**in**	**Toronto**	**yesterday**
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE

Consistent data and annotations

	John	**D.**	**Smith**	**arrived**	**in**	**Toronto**	**yesterday**
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE

Consistent data and annotations

	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE

Consistent data and annotations

	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE
BILUO	B-PER	I-PER	L-PER	O	O	U-LOC	U-DATE

Consistent data and annotations

	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE
BILUO	B-PER	I-PER	L-PER	O	O	U-LOC	U-DATE
BIOES	B-PER	I-PER	E-PER	O	O	S-LOC	S-DATE

Consistent data and annotations

	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE
BILUO	B-PER	I-PER	L-PER	O	O	U-LOC	U-DATE
BIOES	B-PER	I-PER	E-PER	O	O	S-LOC	S-DATE
BMEOW	B-PER	M-PER	E-PER	O	O	W-LOC	W-DATE

Consistent data and annotations

SGML

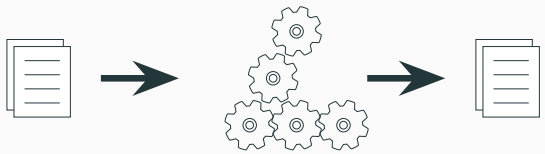
	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE
BILUO	B-PER	I-PER	L-PER	O	O	U-LOC	U-DATE
BIOES	B-PER	I-PER	E-PER	O	O	S-LOC	S-DATE
BMEOW	B-PER	M-PER	E-PER	O	O	W-LOC	W-DATE

Consistent data and annotations

SGML

	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE
BILUO	B-PER	I-PER	L-PER	O	O	U-LOC	U-DATE
BIOES	B-PER	I-PER	E-PER	O	O	S-LOC	S-DATE
BMEOW	B-PER	M-PER	E-PER	O	O	W-LOC	W-DATE

JSON



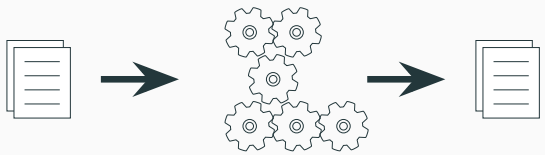
Consistent data and annotations

SGML

	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE
BILUO	B-PER	I-PER	L-PER	O	O	U-LOC	U-DATE
BIOES	B-PER	I-PER	E-PER	O	O	S-LOC	S-DATE
BMEOW	B-PER	M-PER	E-PER	O	O	W-LOC	W-DATE

tabular

JSON



Consistent data and annotations

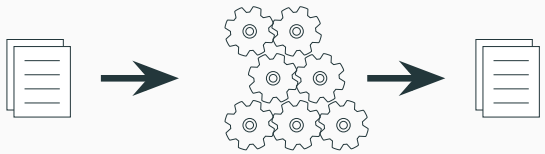
SGML

brat standoff

	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE
BILUO	B-PER	I-PER	L-PER	O	O	U-LOC	U-DATE
BIOES	B-PER	I-PER	E-PER	O	O	S-LOC	S-DATE
BMEOW	B-PER	M-PER	E-PER	O	O	W-LOC	W-DATE

tabular

JSON



Consistent data and annotations

TMX

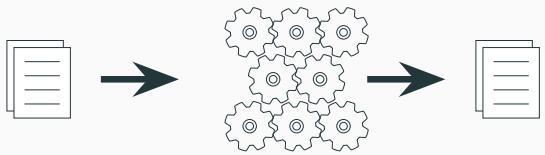
SGML

brat standoff

	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE
BILUO	B-PER	I-PER	L-PER	O	O	U-LOC	U-DATE
BIOES	B-PER	I-PER	E-PER	O	O	S-LOC	S-DATE
BMEOW	B-PER	M-PER	E-PER	O	O	W-LOC	W-DATE

tabular

JSON



Consistent data and annotations

TMX

SGML

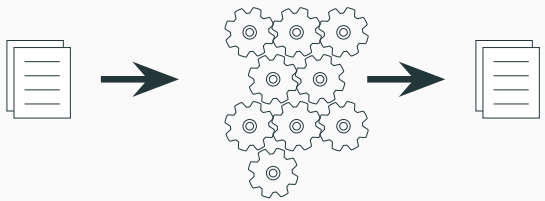
brat standoff

	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE
BILUO	B-PER	I-PER	L-PER	O	O	U-LOC	U-DATE
BIOES	B-PER	I-PER	E-PER	O	O	S-LOC	S-DATE
BMEOW	B-PER	M-PER	E-PER	O	O	W-LOC	W-DATE

tabular

JSON

XML-XCES



Consistent data and annotations

TMX

SGML

brat standoff

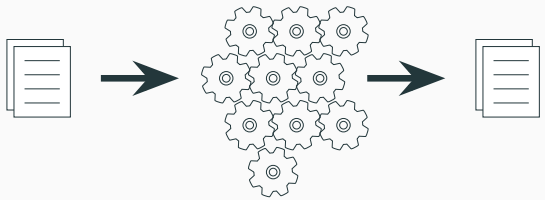
	John	D.	Smith	arrived	in	Toronto	yesterday
BIO	B-PER	I-PER	I-PER	O	O	B-LOC	B-DATE
IOB	I-PER	I-PER	I-PER	O	O	I-LOC	I-DATE
BILUO	B-PER	I-PER	L-PER	O	O	U-LOC	U-DATE
BIOES	B-PER	I-PER	E-PER	O	O	S-LOC	S-DATE
BMEOW	B-PER	M-PER	E-PER	O	O	W-LOC	W-DATE

tabular

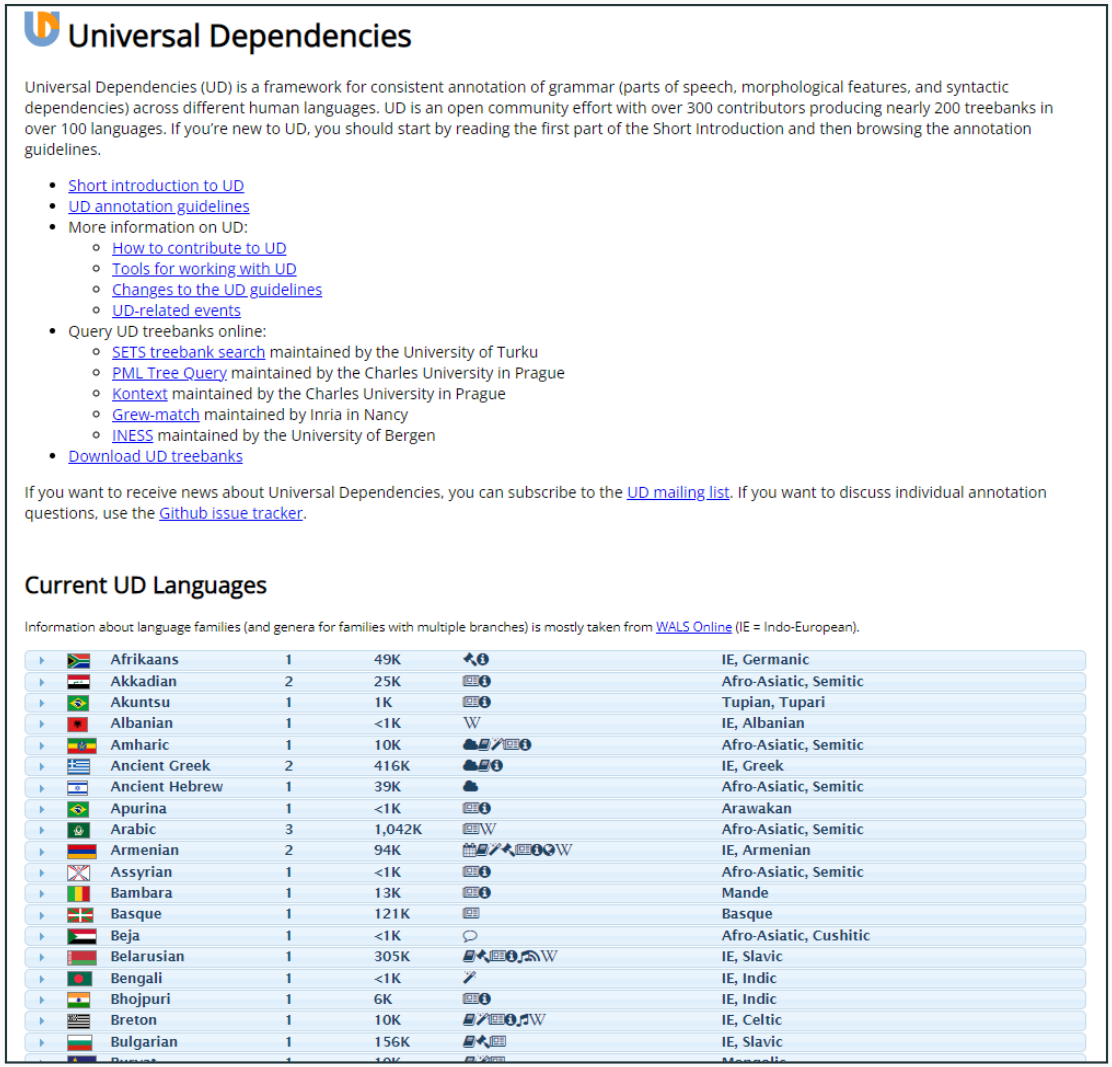
bitext

JSON

XML-XCES



Consistent data and annotations



Universal Dependencies

Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 300 contributors producing nearly 200 treebanks in over 100 languages. If you're new to UD, you should start by reading the first part of the Short Introduction and then browsing the annotation guidelines.

- Short introduction to UD
- UD annotation guidelines
- More information on UD:
  - How to contribute to UD
  - Tools for working with UD
  - Changes to the UD guidelines
  - UD-related events
- Query UD treebanks online:
  - SETS treebank search maintained by the University of Turku
  - PML Tree Query maintained by the Charles University in Prague
  - Kontext maintained by the Charles University in Prague
  - Grew-match maintained by Inria in Nancy
  - INESS maintained by the University of Bergen
- Download UD treebanks

If you want to receive news about Universal Dependencies, you can subscribe to the UD mailing list. If you want to discuss individual annotation questions, use the Github issue tracker.

Current UD Languages

Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

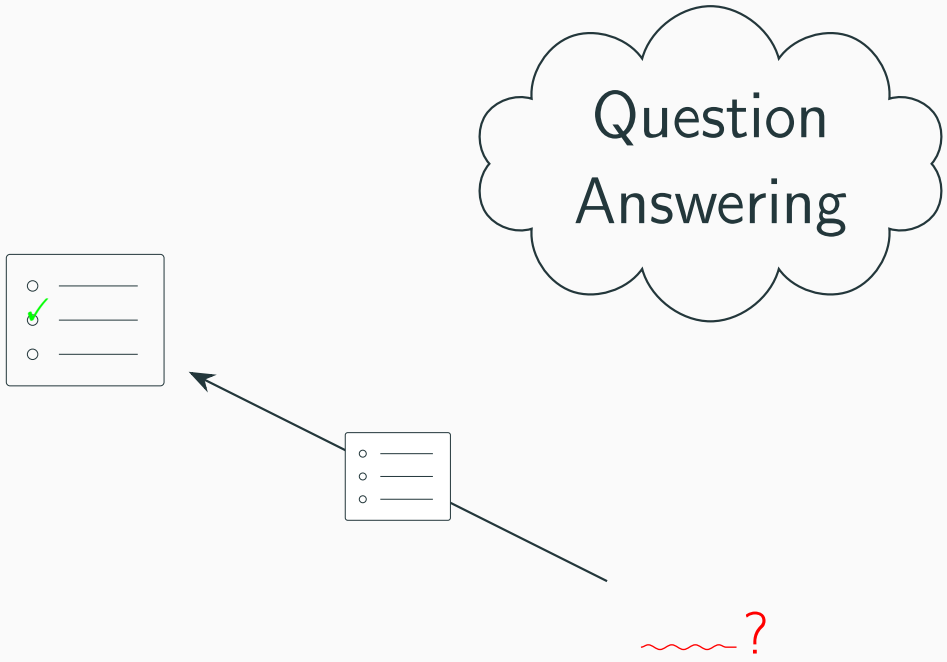
Afrikaans	1	49K		IE, Germanic
Akkadian	2	25K		Afro-Asiatic, Semitic
Akuntsu	1	1K		Tupian, Tupari
Albanian	1	<1K	W	IE, Albanian
Amharic	1	10K		Afro-Asiatic, Semitic
Ancient Greek	2	416K		IE, Greek
Ancient Hebrew	1	39K		Afro-Asiatic, Semitic
Apurina	1	<1K		Arawakan
Arabic	3	1,042K	W	Afro-Asiatic, Semitic
Armenian	2	94K	W	IE, Armenian
Assyrian	1	<1K		Afro-Asiatic, Semitic
Bambara	1	13K		Mande
Basque	1	121K		Basque
Beja	1	<1K		Afro-Asiatic, Cushitic
Belarusian	1	305K	W	IE, Slavic
Bengali	1	<1K		IE, Indic
Bhojpuri	1	6K		IE, Indic
Breton	1	10K	W	IE, Celtic
Bulgarian	1	156K		IE, Slavic

Working on the same task?

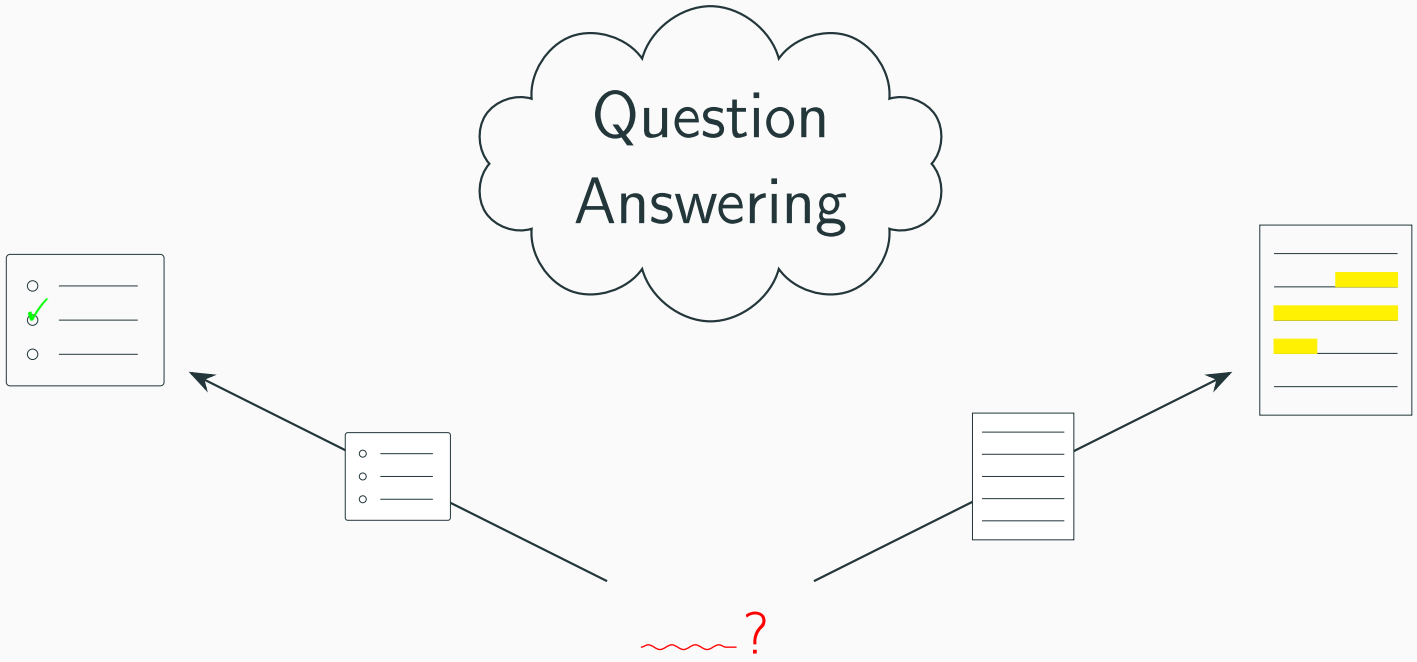
Working on the same task?

?

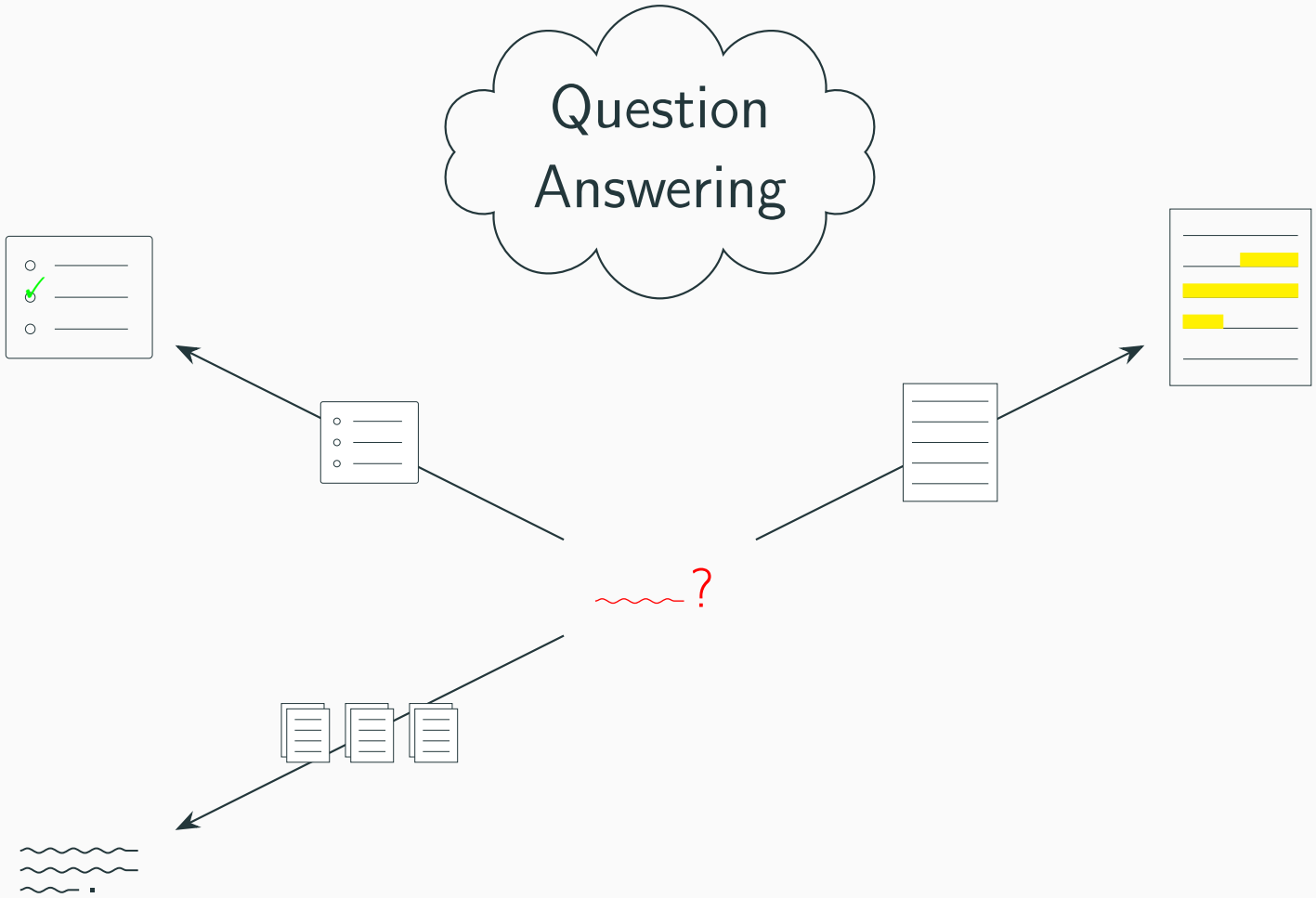
Working on the same task?



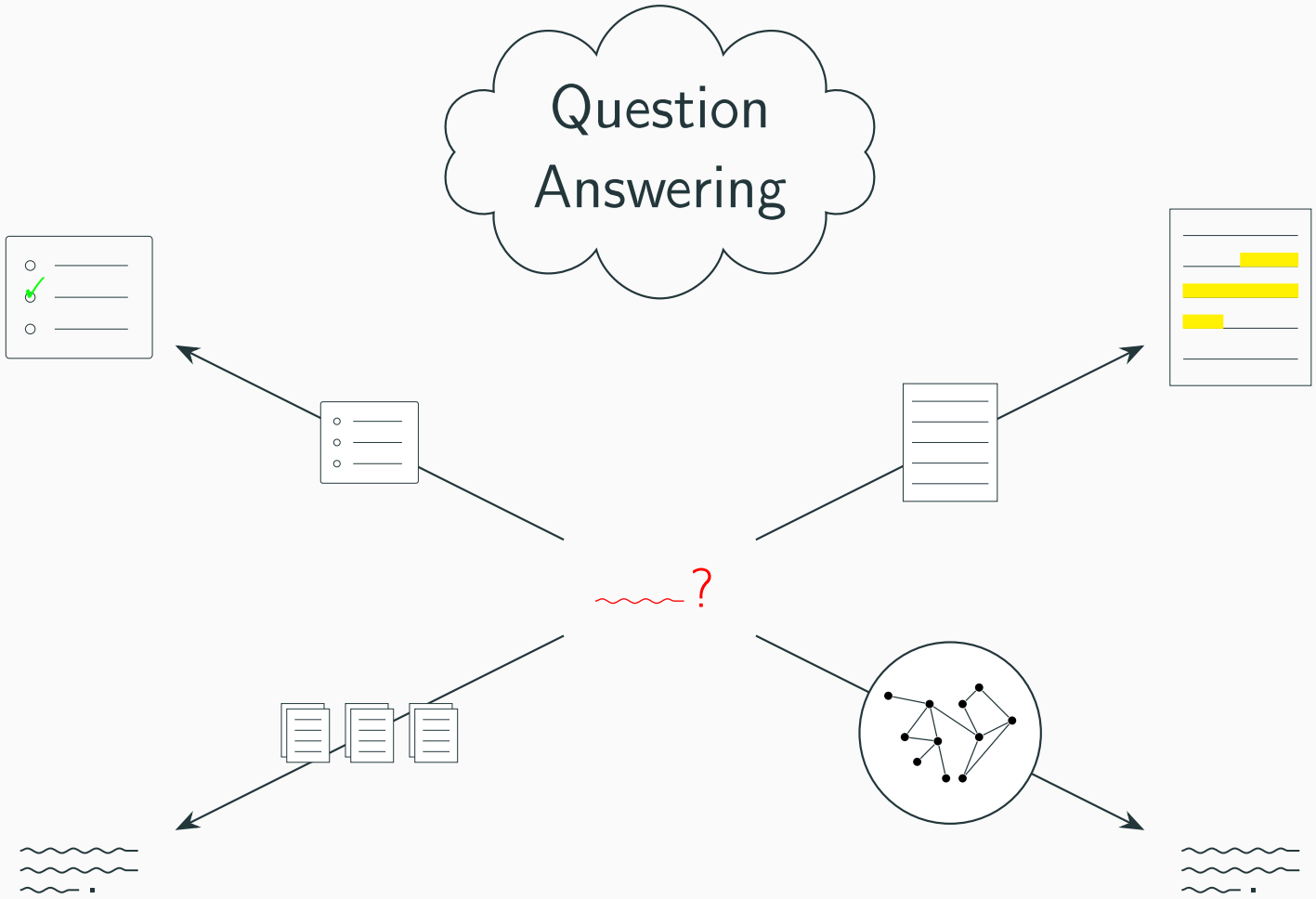
Working on the same task?



Working on the same task?

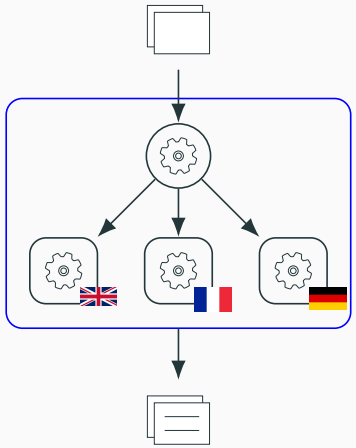


Working on the same task?

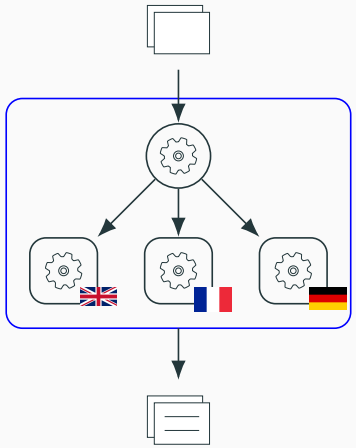


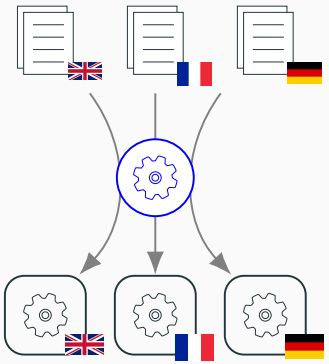
Reflecting upon our work

Reflecting upon our work

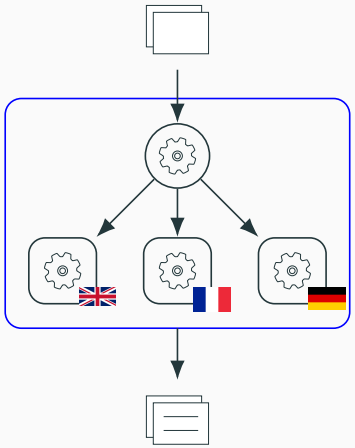


Reflecting upon our work

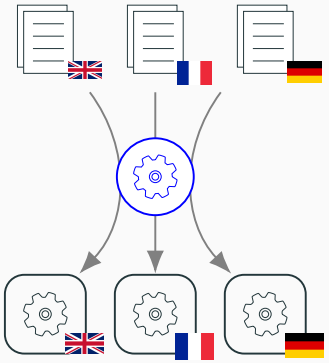


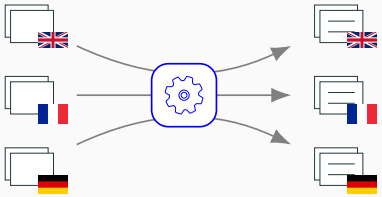


Reflecting upon our work

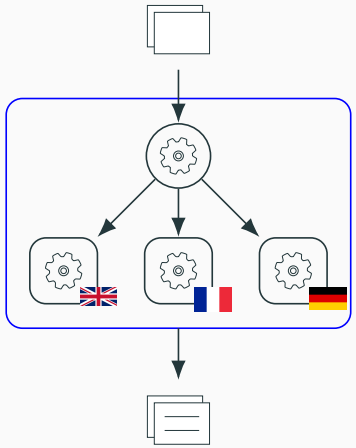


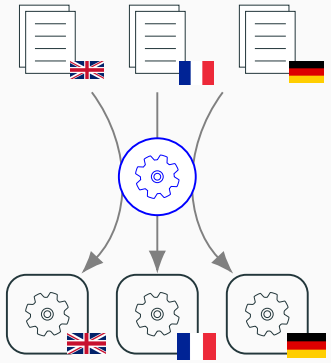
multilingual

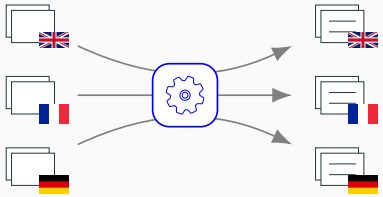




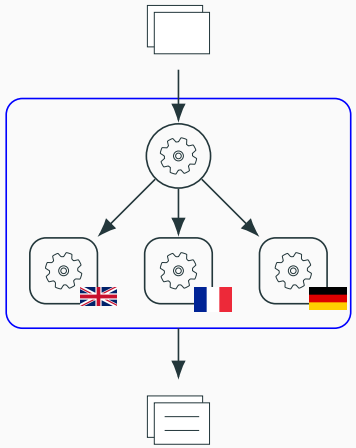
Reflecting upon our work

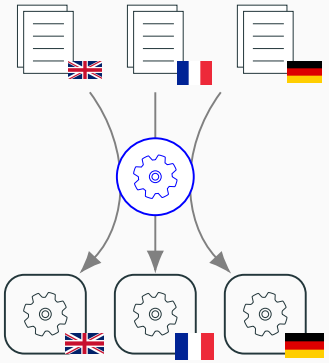


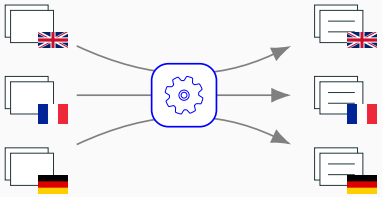




Reflecting upon our work





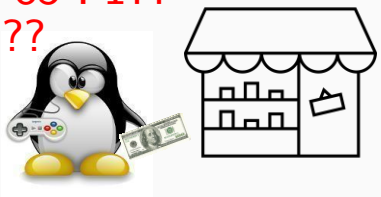


Standards to the rescue!

Benefits and caveats of formal standards

85 F1??

80 F1??



Benefits and caveats of formal standards

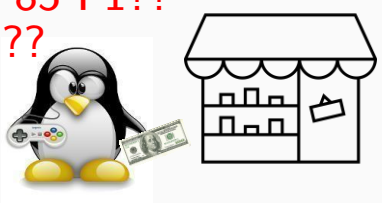
85 F1??

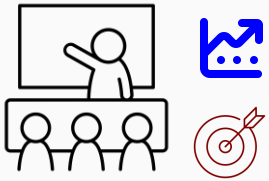
80 F1??

Benefits and caveats of formal standards

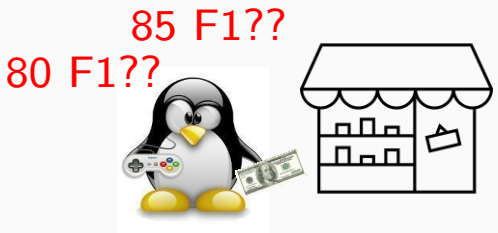
85 F1??

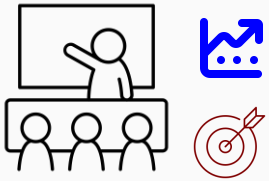
80 F1??

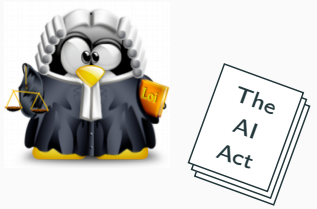




Benefits and caveats of formal standards







Benefits and caveats of formal standards

Key to success: finding an appropriate positioning

- standardize what is consensual, not more, not less
  - ⇝ a forum to build further consensus?
- formalizing/referencing vs prescribing
  - ⇝ preserve research freedom in a fast moving field
- keep up with societal & technological evolutions
  - ⇝ standards need to be revised every few years
- build upon existing work
  - ⇝ promoting existing guidelines rather than recrafting

NLP standardization: current state

NLP standardization: current state

ISO TC 37

"*Language and Terminology*"

NLP standardization: current state

ISO TC 37

"*Language and Terminology*"

ISO IEC JTC 1/SC 42

"*Artificial intelligence*"

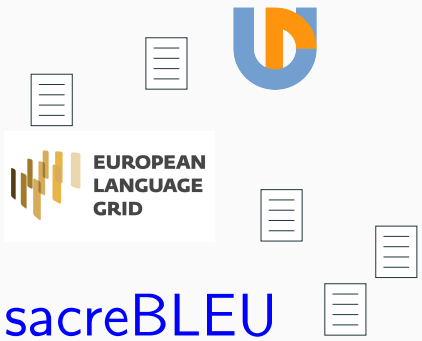
NLP standardization: current state

ISO TC 37

"*Language and Terminology*"

ISO IEC JTC 1/SC 42

"*Artificial intelligence*"



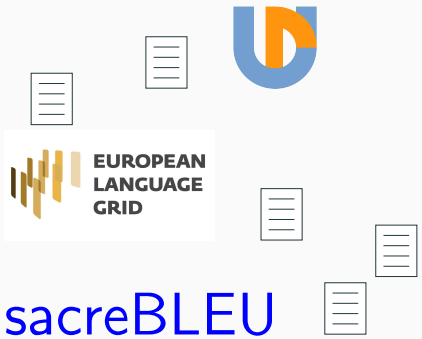
NLP standardization: current state

ISO TC 37

"*Language and Terminology*"

ISO IEC JTC 1/SC 42

"*Artificial intelligence*"



CEN CENELEC JTC 21

"*Artificial intelligence*"

– with NLP projects

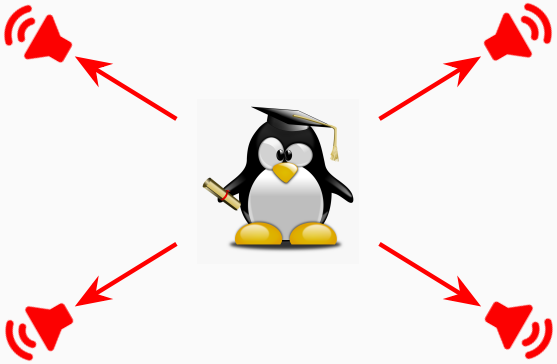
NLP standardization: current state

ISO TC 37

"*Language and Terminology*"

ISO IEC JTC 1/SC 42

"*Artificial intelligence*"



EUROPEAN LANGUAGE GRID

sacreBLEU

CEN CENELEC JTC 21

"*Artificial intelligence*"

– with NLP projects

Take-home messages

- Purposes of standards: **formalizing** common knowledge and guidelines, ensuring **consensus** among experts
- Standardization issues exist **all across NLP**
  ↪ on terminology, concepts, evaluation metrics, data formats...
- A useful contribution to **society at large**
  ↪ expected benefits for researchers, industry, consumers, regulators
- Some good work initiated: guidelines, workshops, projects...
  ↪ next step now is to give them **formal** existence
- Existing groups, to be expanded for better coverage of NLP
  ↪ **researchers needed**: to share your expertise + voice any concern

Thank you for watching!

For any question, comment, suggestion: first.last@inria.fr

And thanks to The Noun Project for the wonderful icons!