Au-delà de l'étiquette unique : Vers une annotation plus riche en TAL

Marie-Catherine de Marneffe

FNRS – UCLouvain – CENTAL

Journées LIFT 2, Orléans, Novembre 2024

Quelle couleur ?

Allez sur **wooclap.com** et utilisez le code **MYCBEC**

Natural Language Inference (NLI)

P: Dana Reeve, la veuve de l'acteur Christopher Reeve, est décédée d'un cancer des poumons à 44 ans.

H: Dana Reeve était atteinte d'une maladie grave.

Entailment – Vrai
Neutral – Non-déterminé ?
Contradiction – Faux

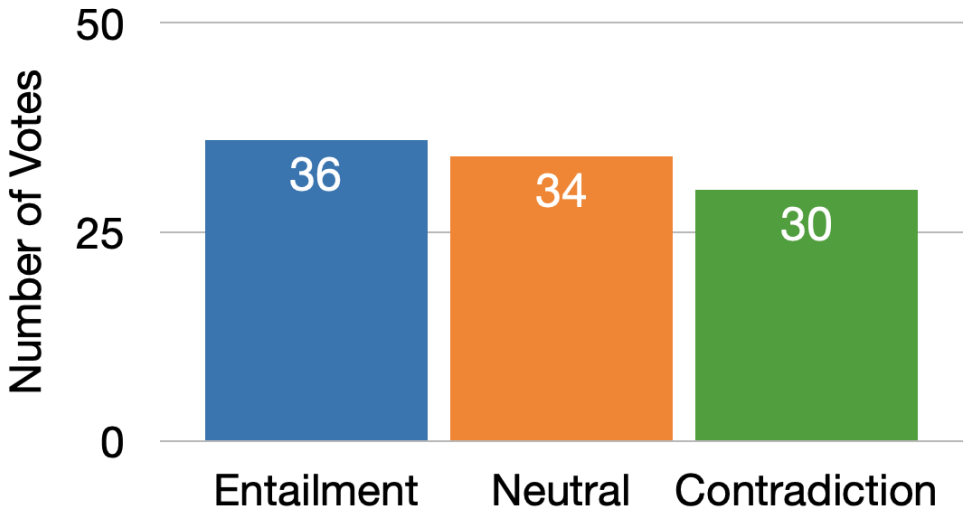
Ici, sur des corpus en anglais

“Human label variation”

[terme de Plank 2022]

P: The park was established in 1935 and was given Corbett's name after India became independent.

H: The park used to be named after Corbett.



[Pavlick and Kwiatowski 2019, Nie et al 2020]

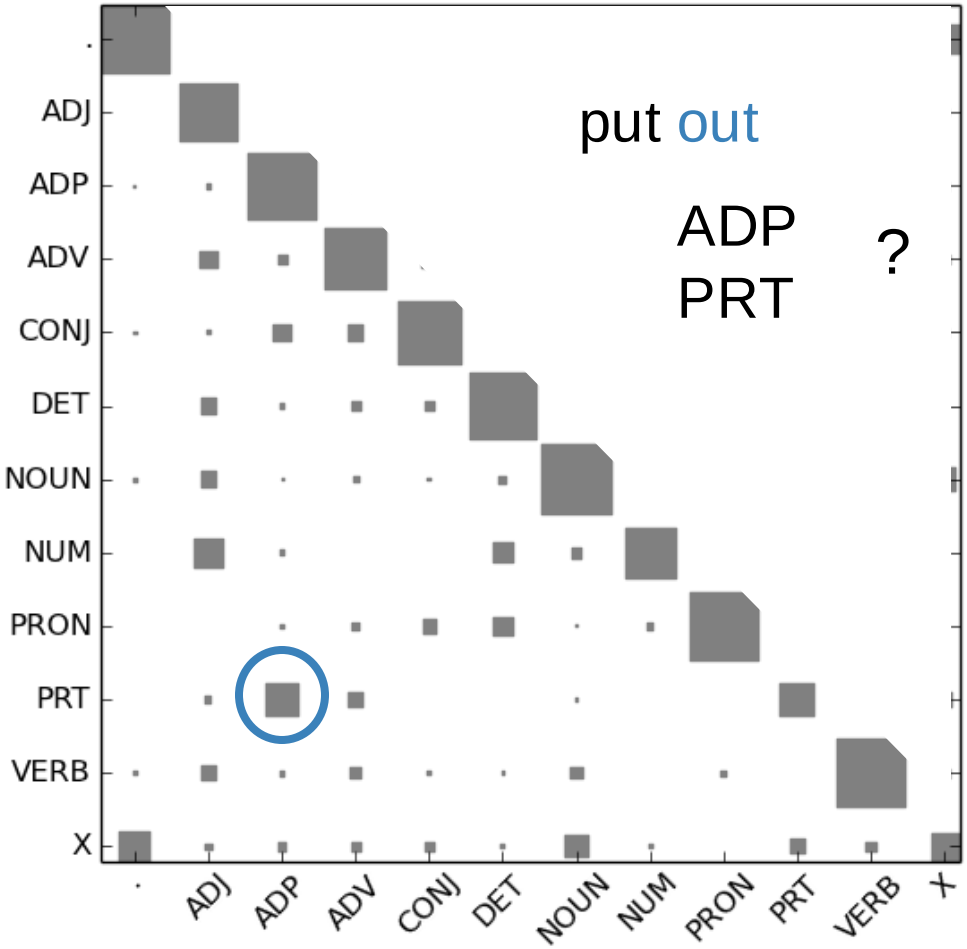
“Truth is a lie: Crowd truth and the 7 myths of human annotation”

[Aroyo & Welty 2015]

1. One truth
2. **Disagreement is bad**
3. Detailed guidelines help
4. **One is enough**
5. Experts are better
6. **All examples are created equal**
7. Once done, forever valid

La variation dans l'annotation est un signal, et non du bruit

[de Marneffe et al. 2012, Aroyo and Welty 2013, Passonneau and Carpenter 2014, Plank et al. 2014, Artstein 2017]



Modéliser la variation d'annotations en NLI

1. Qu'est-ce qui explique cette variation ?
   Quels sont les phénomènes linguistiques ?

2. Peut-on prédire la variation ?

3. Peut-on distinguer la variation intrinsèque de réelles erreurs d'annotation ?

Taxonomie de sources de variation

Probabilistic Enrichment

P: Oh, sorry, wrong church.

H: He or she entered the wrong church.

[E,N,C]: [82, 17, 1]

Coreference

P: Cruises are available from the Bansi Ghat, which is near the City Palace in India.

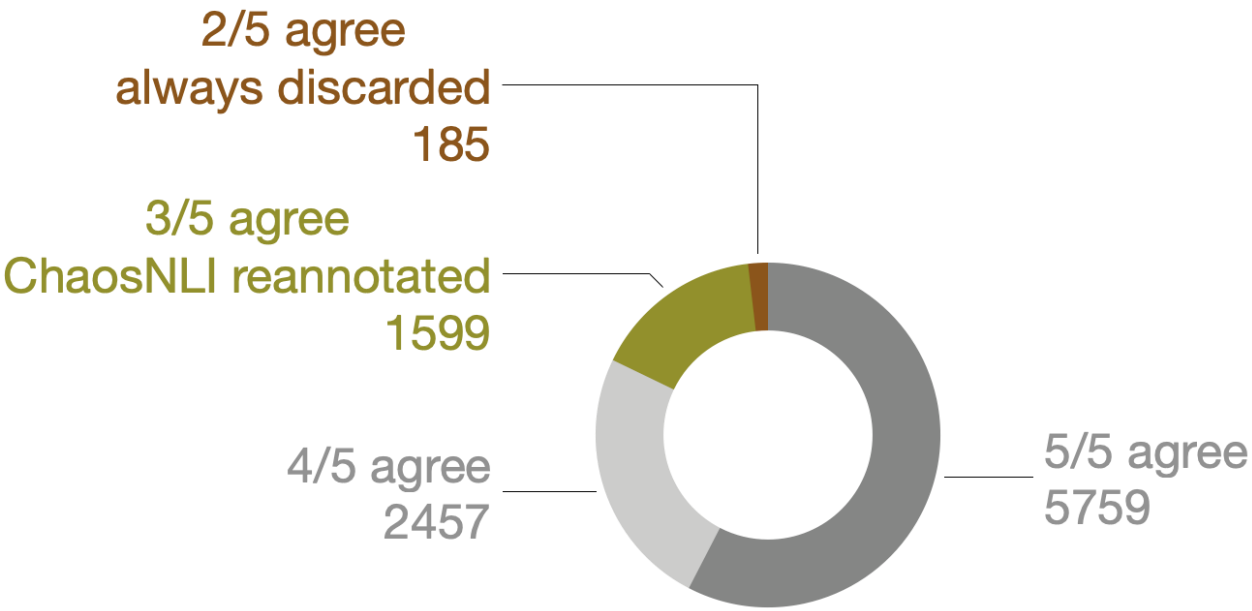
H: You can take cruises from Phoenix, Arizona.

[E,N,C]: [0, 51, 49]

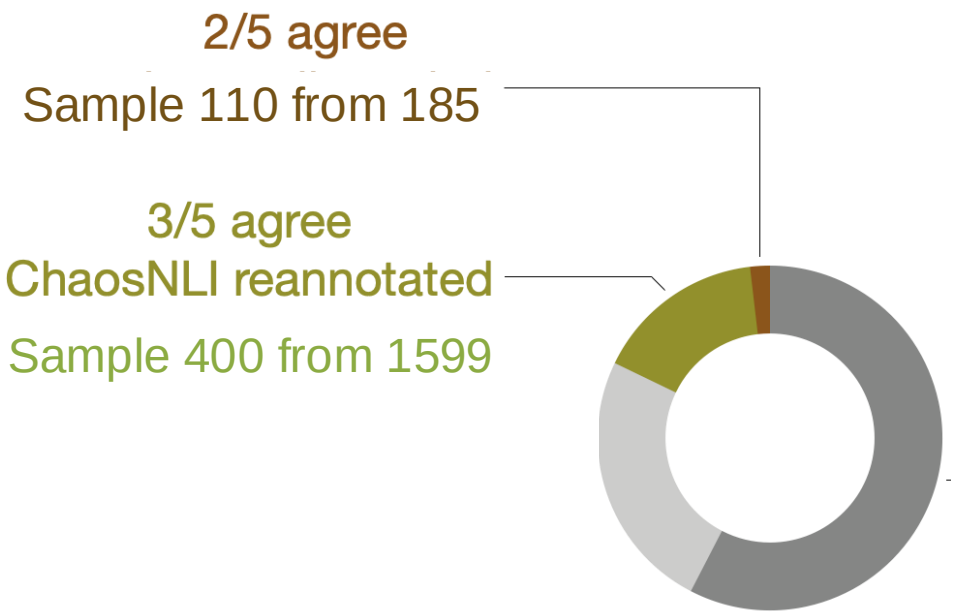
MNLI data dev matched set

10k items, 5 annotations/item [Williams et al. 2018]

ChaosNLI reannotated 3/5 agree, with 100 annotations/item [Nie et al. 2020]

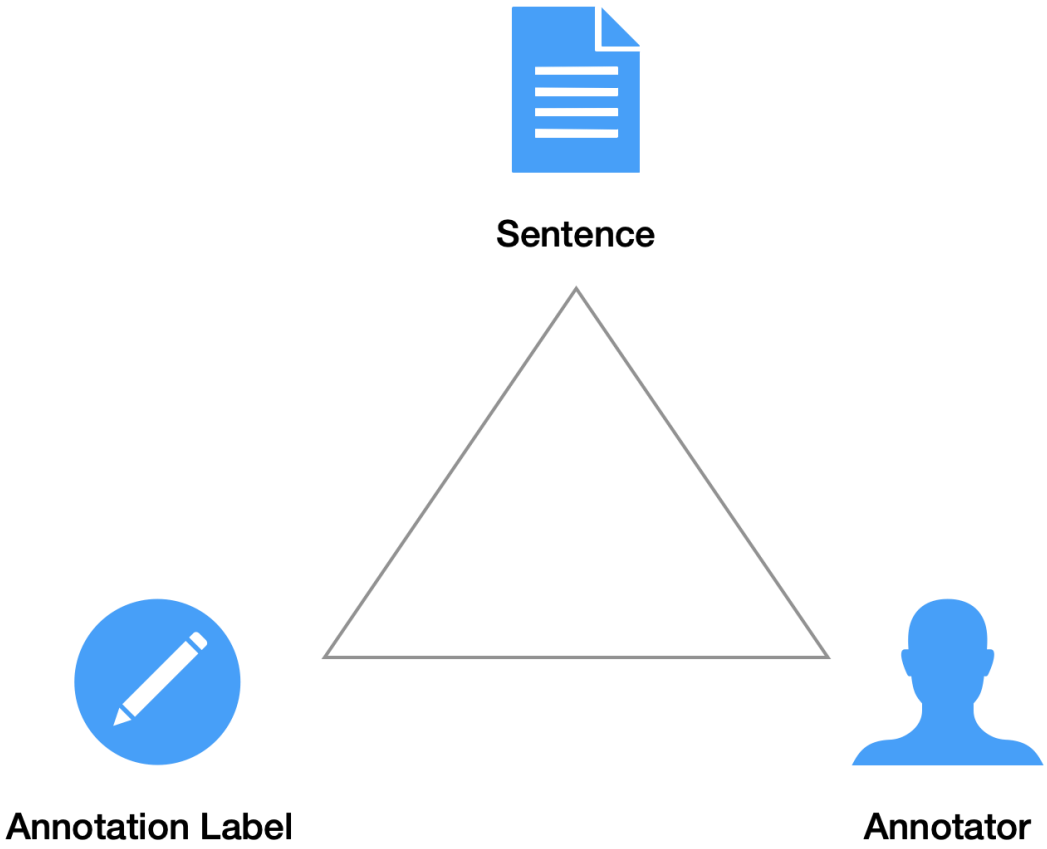


Données pour développer la taxonomie : 510 items

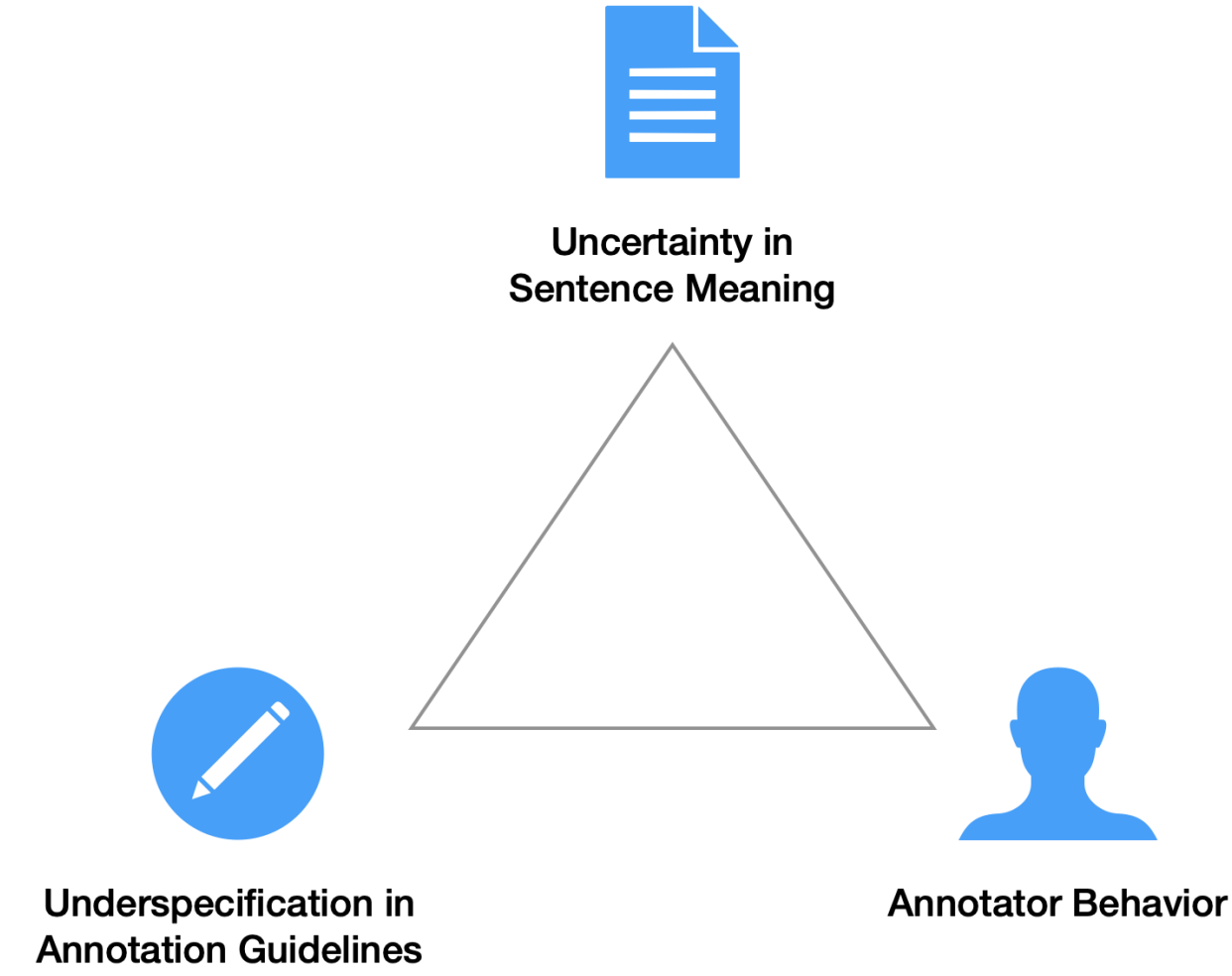


Triangle de référence

[Ogden & Richards 1923, Aroyo & Welty 2013]



Sources de variation dans l'annotation



Processus d’annotation

Round 1
Annotateur #1
a annoté 400 items
— multi-category

développé la taxonomie

écrit le guide d’annotation

Round 2
Annotateurs #1 & #2
ont annoté 110 items

0.69 Krippendorff’s alpha
avec MASI distance

10 catégories pour expliquer la variation

Uncertainty in Sentence Meaning

- Lexical
- Implicature
- Presupposition
- Probabilistic Enrichment
- Imperfection

Underspecification in Annotation Guidelines

- Coreference
- Temporal reference
- Interrogative Hypothesis

Annotator Behavior

- Accommodating minimal underspecified content
- High overlap

Uncertainty in sentence meaning

	Premise	Hypothesis	[E, N, C]
Lexical	**Technological** advances **generally** come in waves that crest and eventually subside.	Advances in **electronics** come in waves.	[82, 17, 1]
Implicature	[…] **some of the most authentic papyrus are sold at The Pharaonic Village** in Cairo […]	The Pharaonic Village in Cairo is the **only** place where one can buy authentic papyrus.	[20, 39, 41]
Presuppos.	What changed?	Nothing changed.	[4, 76, 20]
Probabilistic Enrichment	It's absurd but I can't help it. **Sir James nodded** again.	Sir James thinks it's absurd.	[61, 39, 0]
Imperfection	profit rather	Our profit has not been good.	[3, 90, 7]

Underspecification in guidelines

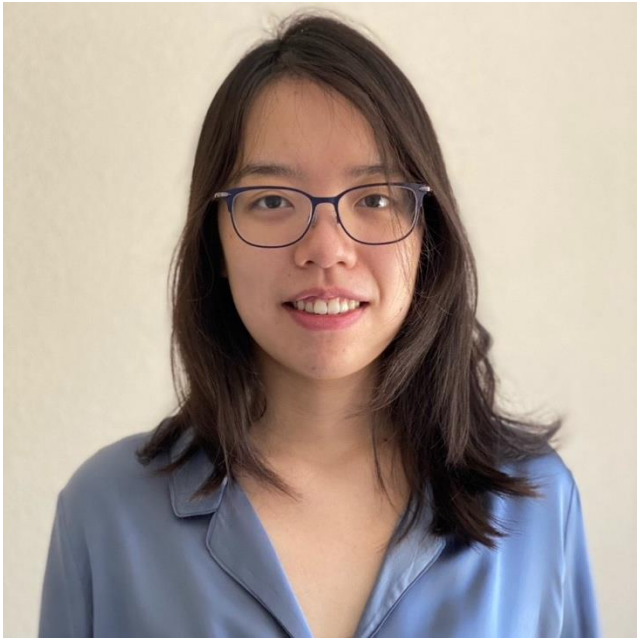
	Premise	Hypothesis	[E, N, C]
Coreference	The original **wax** models of the river gods are on display in the Civic Museum.	They have models made out of **clay**.	[5, 38, 57]
Temporal Reference	However, co-requesters cannot approve additional co-requesters or restrict the timing of the release of the product after it is issued.	They cannot restrict timing of the release of the product.	[90, 8, 2]
Interrogative Hypothesis	was it bad	Was it not good?	[84, 16, 0]

Annotator behavior

	Premise	Hypothesis	[E, N, C]
Accommodating minimal underspecified content	Indeed, 58 percent of Columbia/HCA's beds lie empty, compared with 35 percent of nonprofit beds.	58% of Columbia/HCA's beds are empty, **said the report**.	[97, 3, 0]
	After four years, Clinton has learned how to avoid looking unpresidential.	After four **torturous** years, Clinton finally gets how to avoid unpresidential behavior.	[49, 48, 3]
High overlap	Yet, in the mouths of the white townsfolk of Salisbury, N.C., it sounds convincing.	White townsfolk in Salisbury, N.C. think it sounds convincing.	[68, 27, 5]

Investigating reasons for disagreement in NLI

Nan-Jiang Jiang & MC de Marneffe TACL 2022



Certaines sources mènent-elles à plus d'accord?

	Converge (%)	# items
Lexical	17.7	124
Implicature	12.5	24
Presupposition	0.0	12
Probabilistic Enrichment	13.3	165
Imperfection	22.7	22
Coreference	14.7	75
Temporal Reference	25.0	12
Interrogative Hypothesis	20.0	15
Accommodating content	25.5	51
High Overlap	0.0	8

Certaines sources mènent-elles à plus d'accord?

	Converge (%)	# items
Lexical	17.7	124
Implicature	12.5	24
Presupposition	0.0	12
Probabilistic Enrichment	13.3	165
Imperfection	22.7	22
Coreference	14.7	75
Temporal Reference	25.0	12
Interrogative Hypothesis	20.0	15
Accommodating content	**25.5**	**51**
High Overlap	0.0	8

Types différents de texte ajouté

P: Indeed, 58% of Columbia/HCA's beds lie empty, compared […]

H: 58% of Columbia/HCA's beds are empty, **said the report**.

[E,N,C]: [**97**, 3, 0]

P: And here, current history adds a major point.

H: Current **American** history adds a major point.

[E,N,C]: [**32**, 67, 1]

P: The equipment you need for windsurfing can be hired from the beaches at Faro.

H: Windsurfing equipment is available for hire in Faro **all year round**.

[E,N,C]: [**7**, 93, 0]

[Potts 2005, Simons et al. 2010, McNally 2016]

3 approches pour modéliser la variation

Entraîner sur la distribution et prédire une distribution

4-way classification: E, N, C, Complicated

[Kenyon-Dean et al. 2015]

Multi-label classification: one or more of E, N, C

[i.a., Passonneau et al. 2012]

.

On se concentre sur 2 approches

4-way classification: E, N, C, Complicated

[Kenyon-Dean et al. 2015]

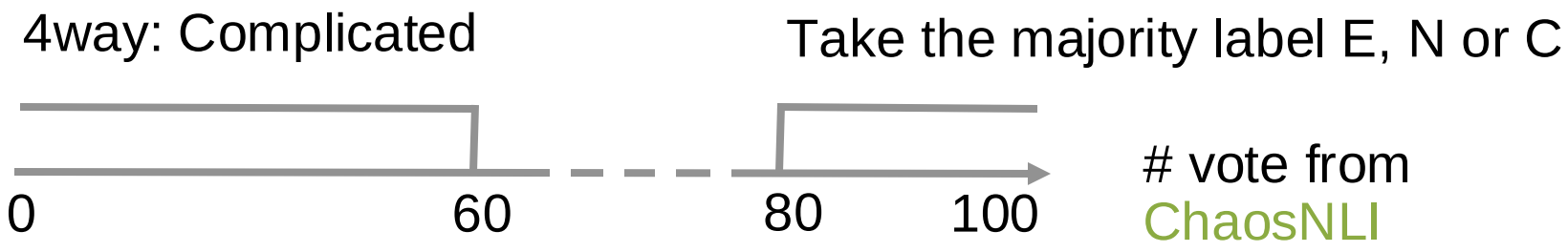
Multi-label classification: one or more of E, N, C

[i.a., Passonneau et al. 2012]

Peaufiner (fine-tune) RoBERTa

Baseline MixUp [Zhang et al. 2021]
entraîne sur la distribution, puis convertit vers une étiquette

Comment assigner les étiquettes ?



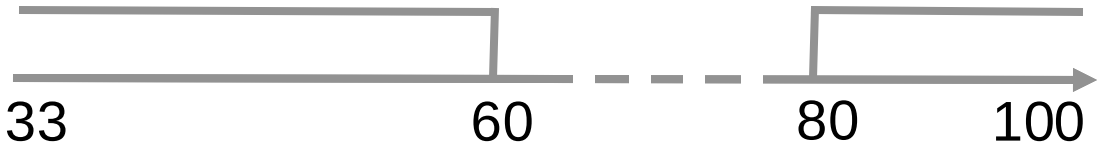
	E	N	C	
Chaos	195	57	37	604 Complicated

Comment assigner les étiquettes ?

Multi-label: any label with >20 votes

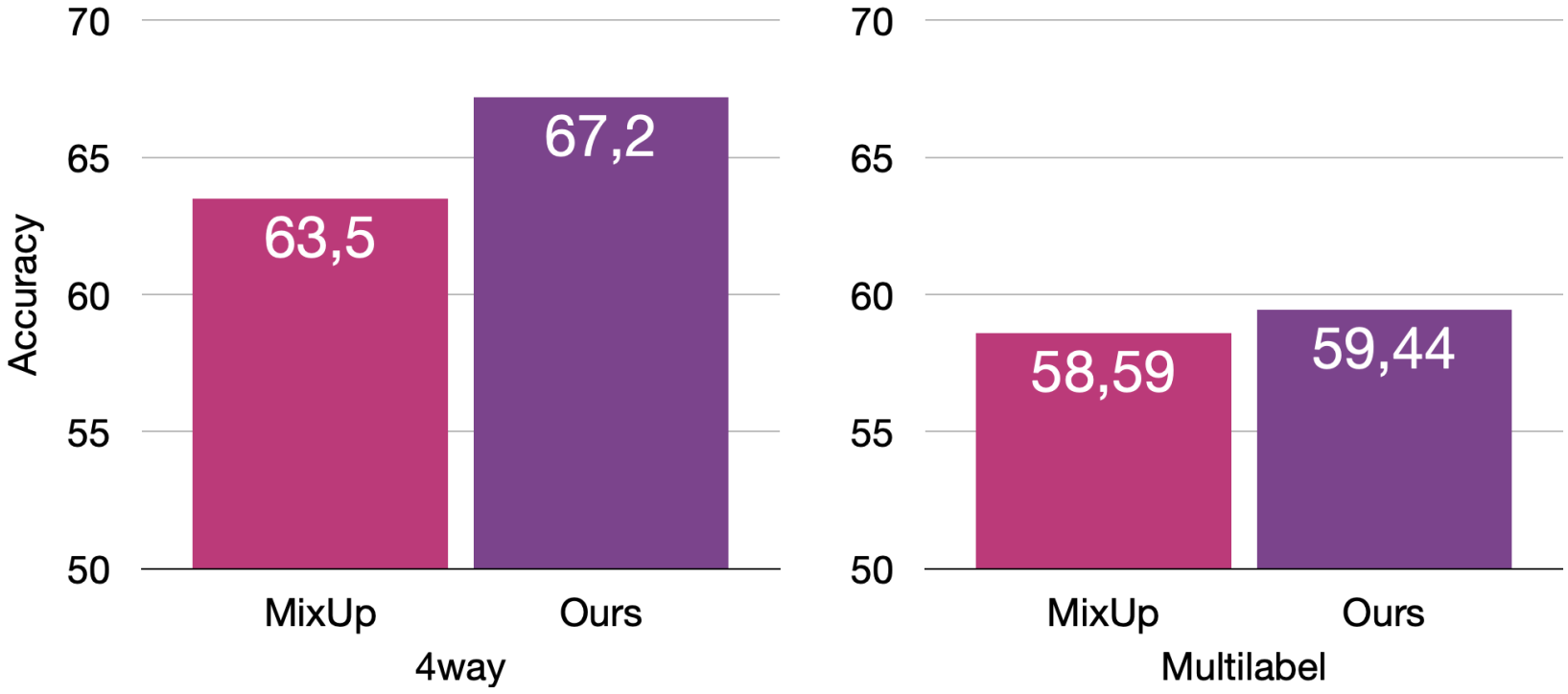
4way: Complicated

Take the majority label E, N or C



	E	N	C	EN	NC	EC	ENC
Chaos	195	57	37	291	205	32	76
				604 Complicated			

Notre modèle est meilleur que la baseline



Plus d'information avec le multi-label

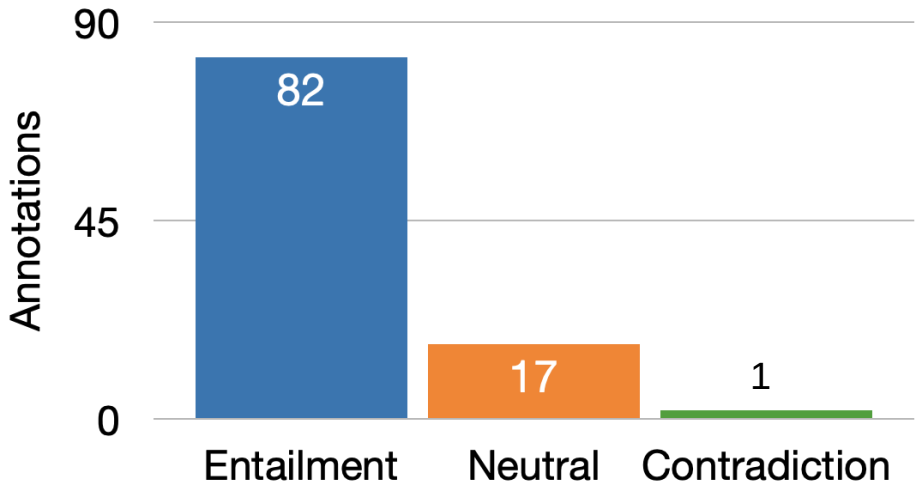
Probabilistic Enrichment

P: Oh, sorry, wrong church.

H: He or she entered the wrong church.

4way prediction: E

Multilabel: EN



Gold (w/ threshold): E

Comment distinguer les vraies erreurs ?

VariErr NLI: Separating annotation error from human label variation

Leon Weber-Genzel, Siyao Peng, MC de Marneffe & Barbara Plank, ACL 2024

VariErr : Annotations en deux passes

P: Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.

H: Newspaper preprints can cost as much as $5.

1er round : étiquette et explication

P: Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.

H: Newspaper preprints can cost as much as $5.

Round 1: NLI Label & Explanation

L	Explanation
E	5 dollars for a piece of newspaper.
N	The context only mentions how low the price may be, not how high it may be. The maximum cost of newspaper preprints is not given in the context.
C	The context says 5 or 6 cents, not $5.

500 ChaosNLI items, 1 933 paires étiquette-explication

2e round : validation

P: Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.

H: Newspaper preprints can cost as much as $5.

	Round 1: NLI Label & Explanation	*Round 2: Validity*			
L	Explanation	1	2	3	4
E	5 dollars for a piece of newspaper.	✗	✗	✗	✗
N	The context only mentions how low the price may be, not how high it may be.	✓	✓	✓	✓
	The maximum cost of newspaper preprints is not given in the context.	✓	✓	✓	✓
C	The context says 5 or 6 cents, not $5.	✗	✗	✓	✓

500 ChaosNLI items, 1 933 paires étiquette-explication

Définition stricte d’erreur

VariErr : identification des erreurs

P: Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.

H: Newspaper preprints can cost as much as $5.

		Round 1: NLI Label & Explanation	*Round 2: Validity*			
L	A	Explanation	1	2	3	4
E	4	5 dollars for a piece of newspaper.	✗	✗	✗	✗
N	1	The context only mentions how low the price may be, not how high it may be.	✓	✓	✓	✓
	3	The maximum cost of newspaper preprints is not given in the context.	✓	✓	✓	✓
C	2	The context says 5 or 6 cents, not $5.	✗	✗	✓	✓

self-validation

500 ChaosNLI items, 1 933 paires étiquette-explication

Définition stricte d'erreur

88.57% self-validées, 82.82% peer-validées

37% des items pour lesquels il y avait une erreur d'étiquette

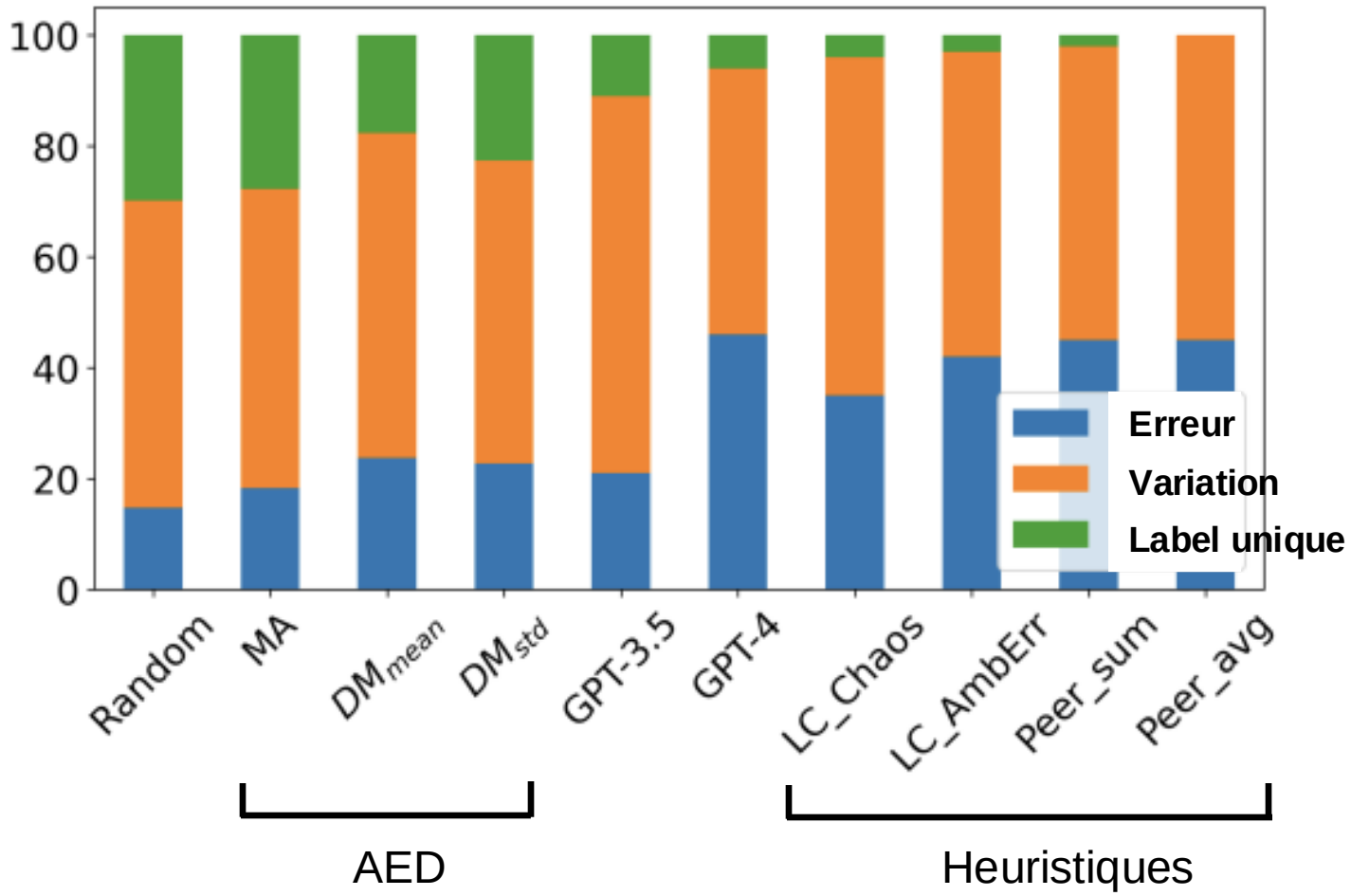
Identifier les erreurs automatiquement ?

Detection automatique d'erreurs (AED) comme une tâche de ranking

Comparaison des listes rankées aux erreurs

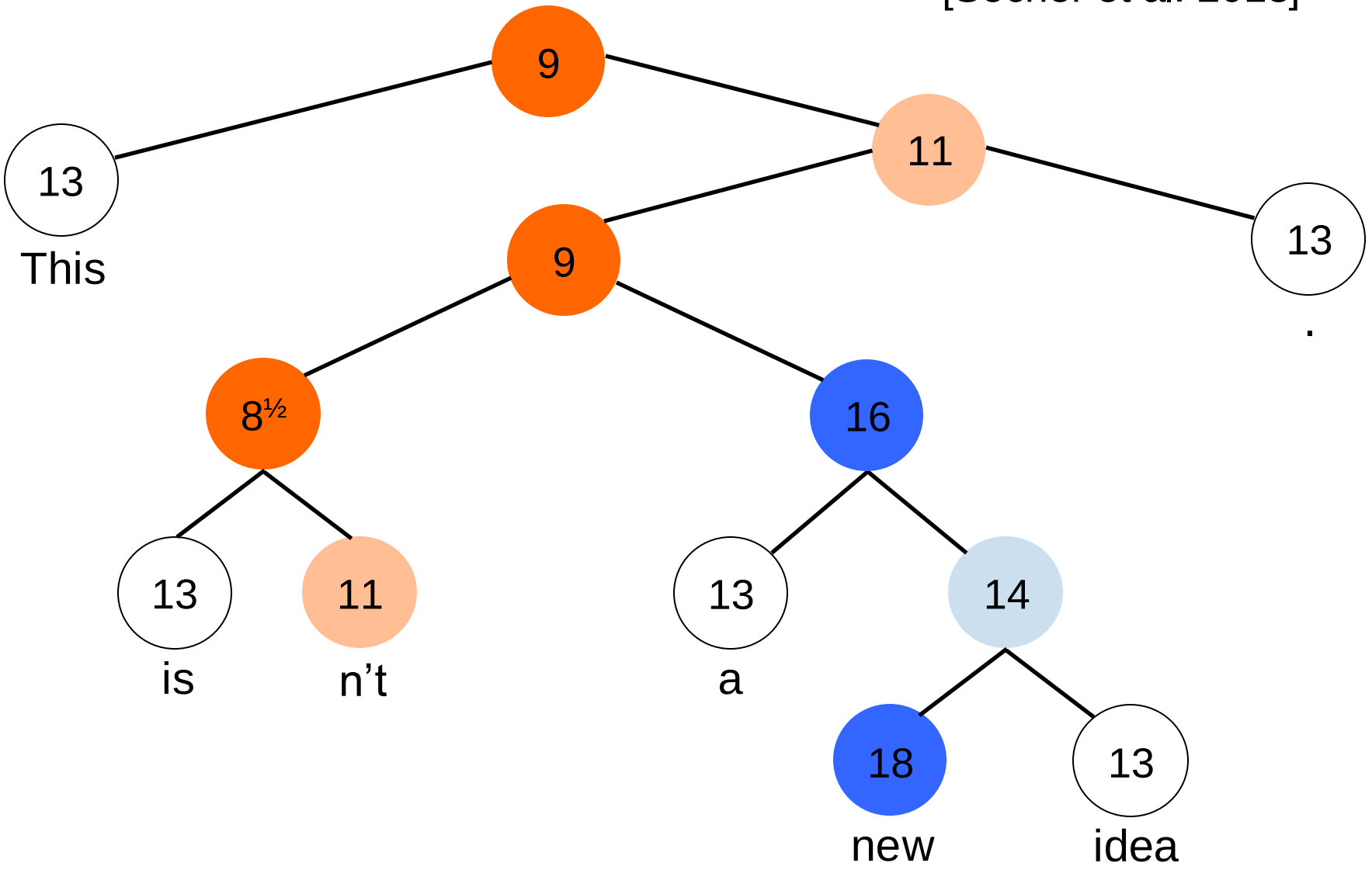
	Exactitude
3 systèmes AED d'état de l'art	22.8
GPT-4	31.3
Heuristiques sur ChaosNLI/VariErr	32.5 - 46.5

Proportion de types d'items dans le top 100



Stanford Sentiment Treebank

[Socher et al. 2013]



Phrases qui ont un score de 0.5

a film as Byatt fans could hope for	13,15,15
A sly dissection of the inanities of the contemporary music business and a rather sad story of the difficulties	13,15,15

Phrases qui ont un score de 0.5

a film as Byatt fans could hope for	13,15,15
A sly dissection of the inanities of the contemporary music business and a rather sad story of the difficulties	13,15,15

---

a sad, superior human	9,13, 21
a joke	8,15, 20
Angel presents events partly from the perspective of Aurelie and Christelle and infuses the film with the sensibility of a particularly nightmarish fairytale	9,17 17

Accepter la variation d’étiquettes

Publier toutes les annotations !

Pour permettre une compréhension robuste du langage naturel, les modèles doivent distinguer les éléments sur lesquels les êtres humains sont le plus souvent d'accord de ceux qui donnent lieu à de la variation.

Merci !

TACL action editor Anette Frank, the anonymous reviewers, Micha Elsner, Michael White, for insightful feedback, and Angélica Aviles Bosques for her help with the annotations.

github.com/njjiang/NLI_disagreement_taxonomy
github.com/mainlp/VariErr-NLP