PhD Defense

Extraction and normalization of simple and structured entities in medical documents

PhD Student: **Perceval Wajsburt** [1]

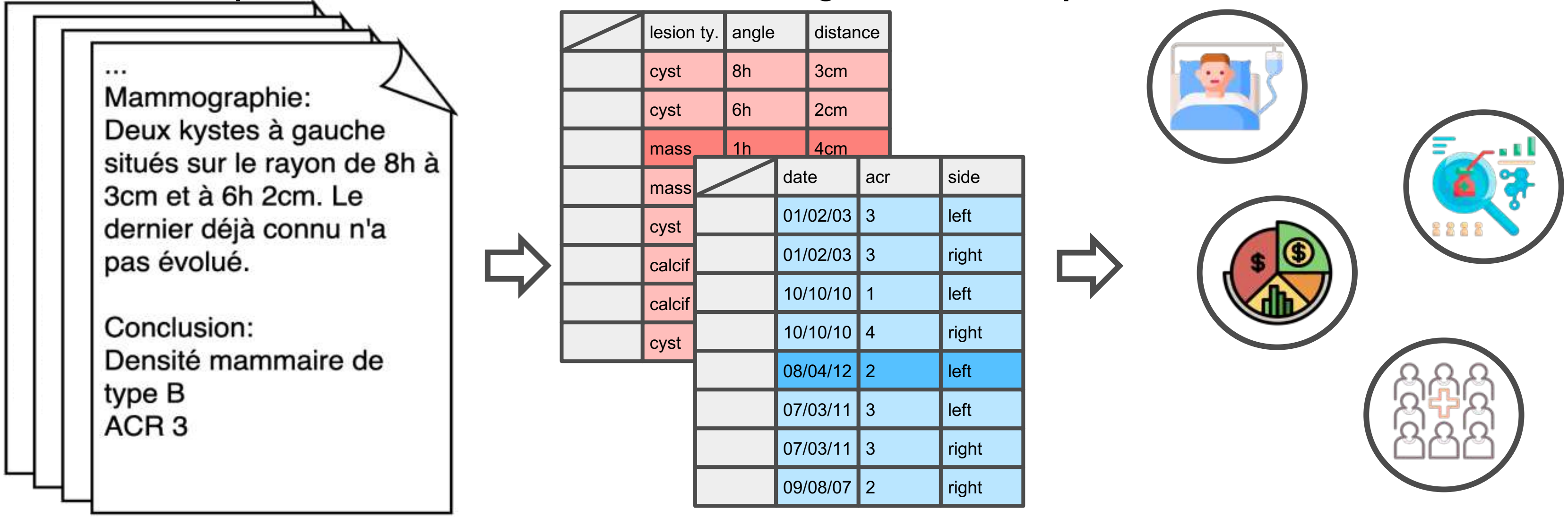
Supervisors: **Xavier Tannier**[1], **Christel Daniel** [2]

[1, 2] Sorbonne Université, LIMICS || {*perceval.wajsburt, xavier.tannier*}*@sorbonne-universite.fr*
[2] Assistance Publique des Hôpitaux de Paris, LIMICS, || {christel.daniel@aphp.fr}

Introduction

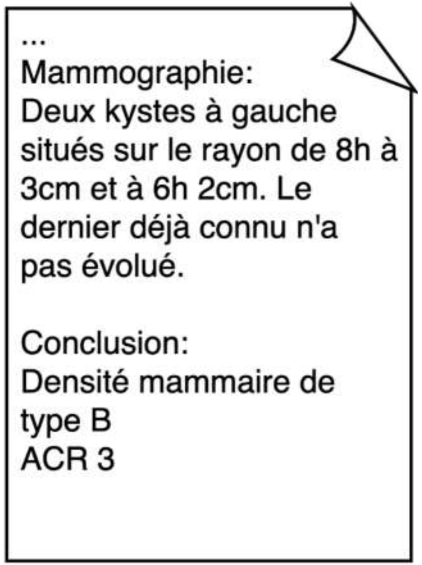
Information extraction

- ➔ A large quantity of information in **textual format**
- ➔ Medical research, patient care, hospital management need **structured data**
- ➔ to follow patients, build cohorts, manage services, perform statistical studies...



Information extraction

- ➔ To fill these structured databases, we **extract information** from documents
- ➔ Depending on the need, we extract more or less complex entities

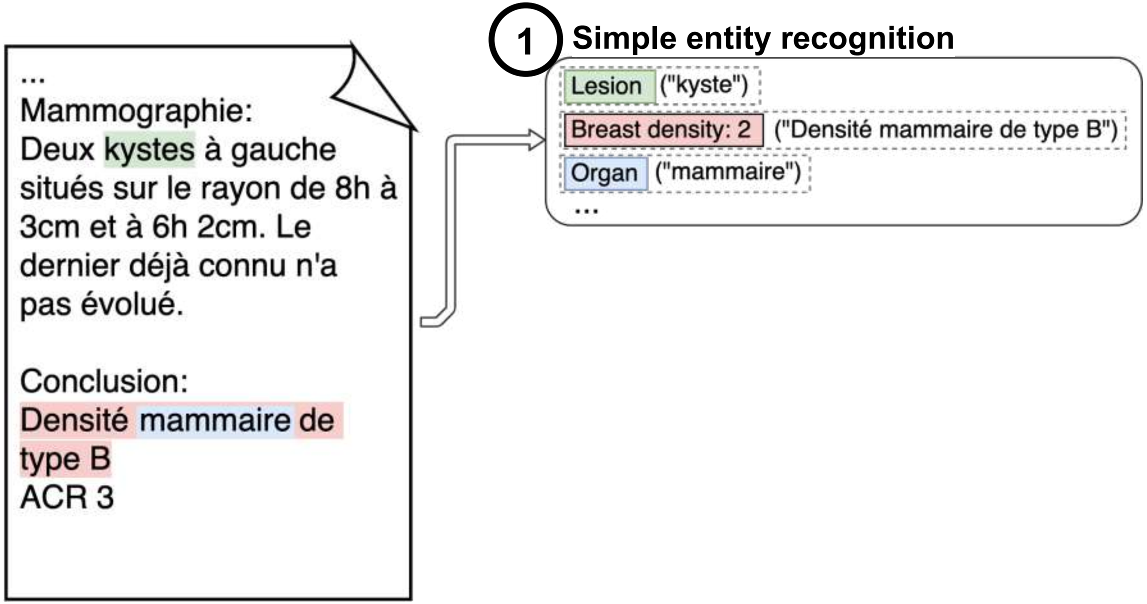


...
Mammographie:
Deux kystes à gauche situés sur le rayon de 8h à 3cm et à 6h 2cm. Le dernier déjà connu n'a pas évolué.

Conclusion:
Densité mammaire de type B
ACR 3

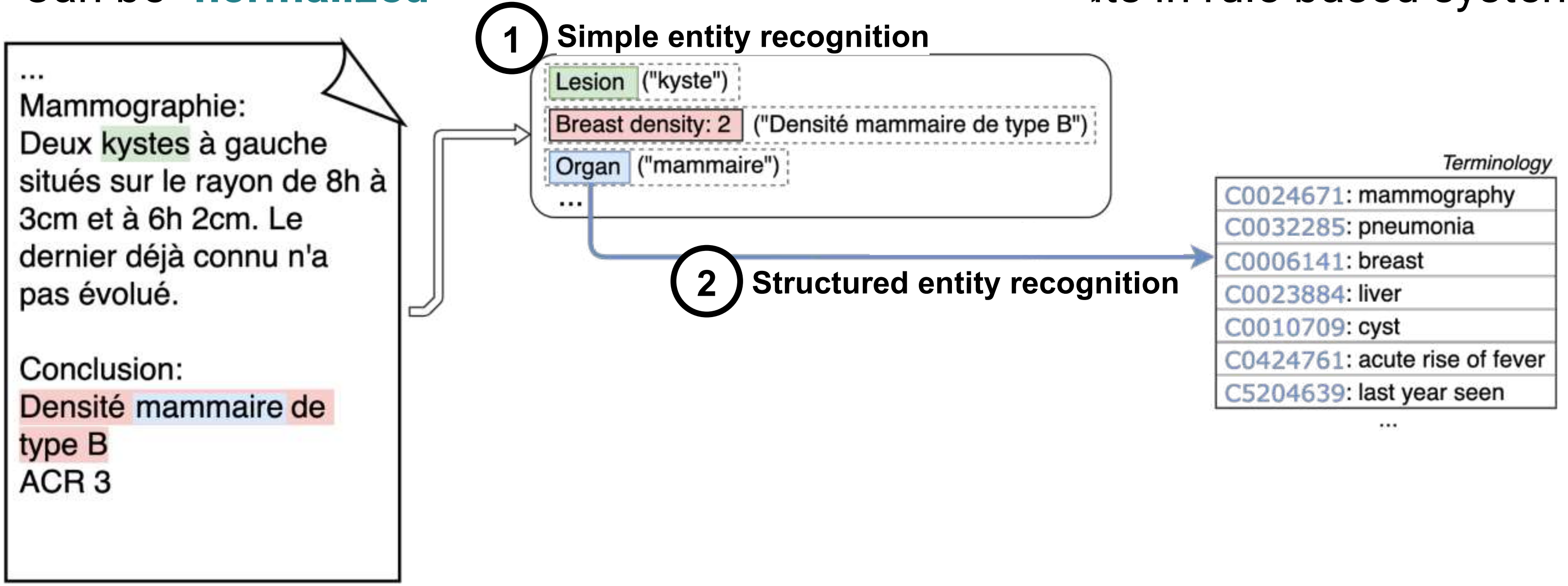
Information extraction

➔ "**Named entities**" used as is, or as building bricks for more complex objects 1



Information extraction

- ➔ “**Named entities**” used as is, or as building bricks for more complex objects 1
- ➔ Can be “**normalized**” to be queriable, or used as inputs in rule-based systems 2

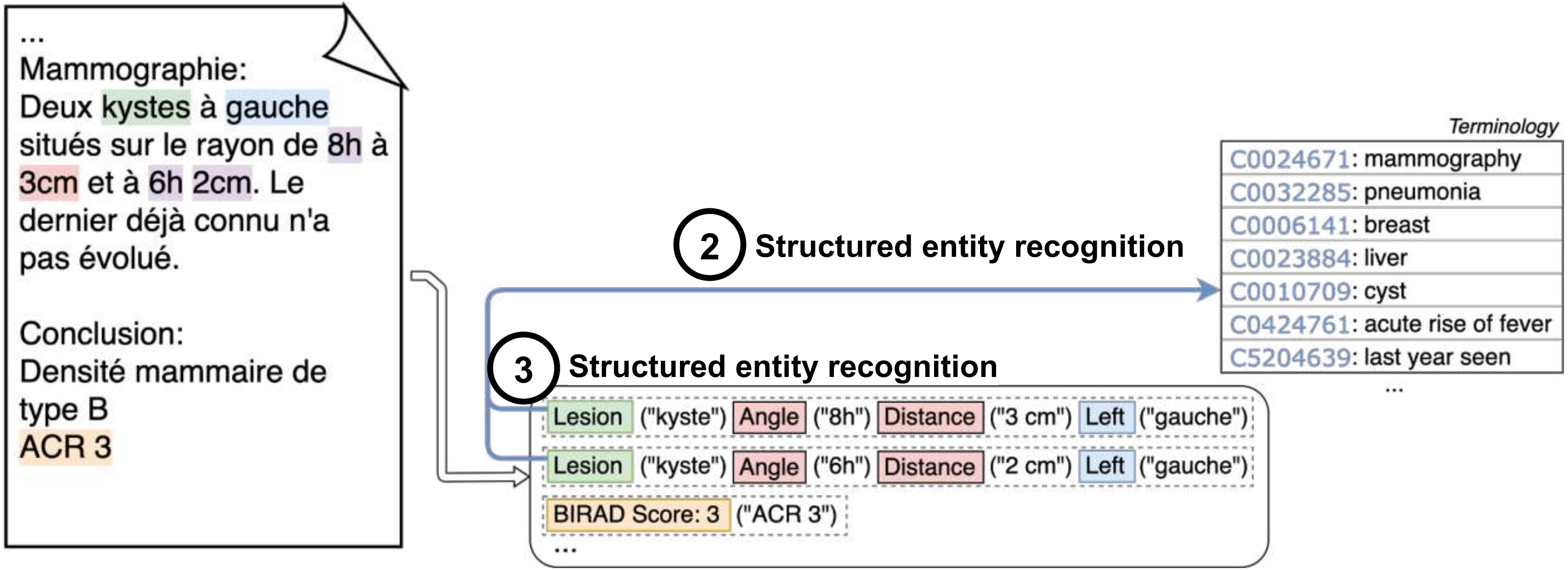


Information extraction

→ For some needs, normalized named entities are not enough

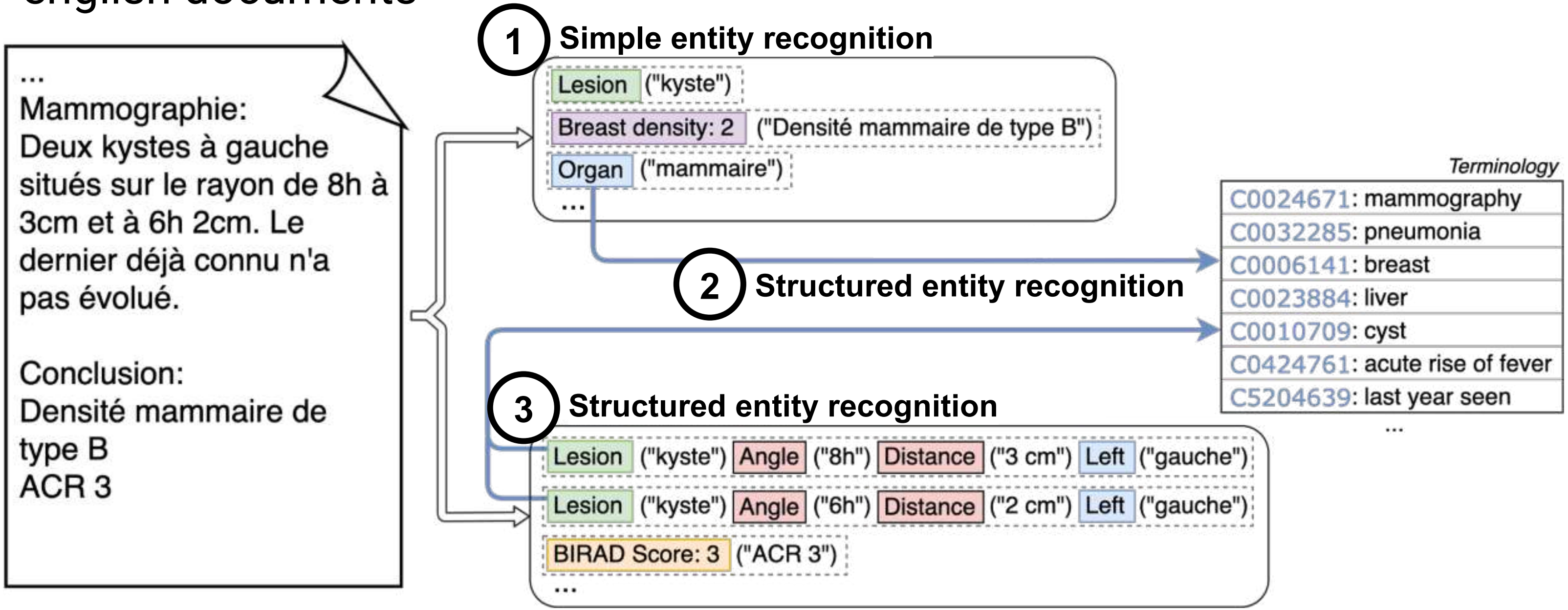
→ We extract **structured** entities, with multiple parts, and multiple labels

3



Information extraction

➔ We tackle these **three** tasks in the biomedical and clinical domain, with non-english documents



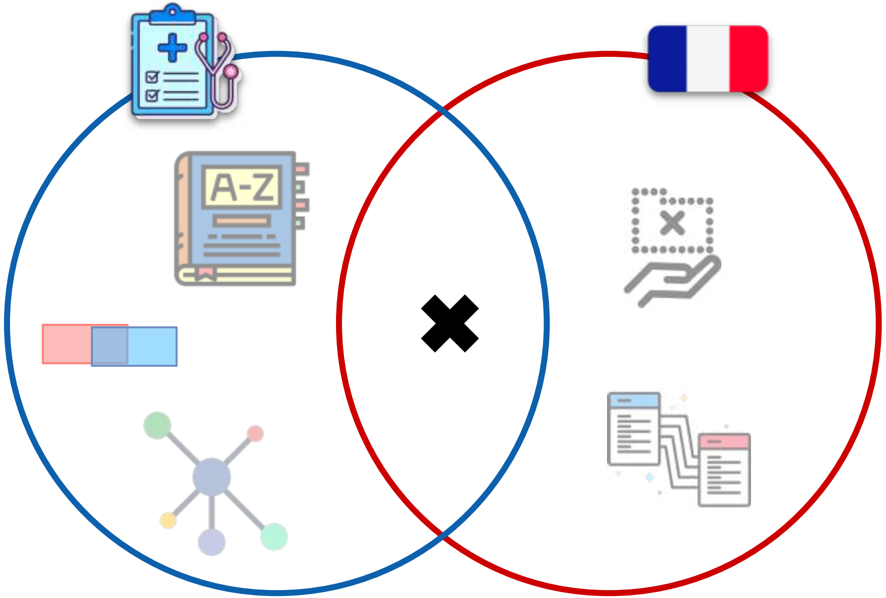
Domain

Medical domain

- ➔ Technical vocabulary
- ➔ Inherent semantic structure
- ➔ Overlapping information
- ➔ Textual structure (sections, lists...)

French domain

- ➔ Fewer available resources and tools
- ➔ Less research literature
- ➔ Need to be interoperable with widespread mainly English resources like terminologies & ontologies



Machine learning

Pros

- ➔ Automatic feature extraction and capture hidden patterns
- ➔ Better generalization (especially with pre-training e.g. BERT)
- ➔ Improvement through sample correction (i.e. more data)

Cons

- ➔ Requires complex architectures and lots of samples
- ➔ Hard to interpret: “blackboxes”

Research questions

- ➔ How do we cope with **overlapping** information in textual documents ?
- ➔ In **low-data** contexts, can we leverage resources in **other languages** or existing **medical knowledge** to bootstrap and improve our models ?
- ➔ How do we **represent** information to extract **complex entities** from clinical reports ?
- ➔ How do we extract structured entities composed of **different parts & labels** ?

Nested named entity recognition

Classic named entity recognition

- ➔ A named entity is a typed span of text
- ➔ Classically, we predict a tag for each word (e.g. using the BIOUL scheme)
- ➔ Example with two named entities in a sentence:

La patiente a un petit nodule dans le quadrant sup. ext. du sein droit .

(= The patient has a small nodule in the upper outer quadrant of the right breast.)

La	patiente	a	un	petit	nodule	dans	le	quadrant	sup.	ext.	du	sein	droit	.
⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩
O	O	O	O	O	U-lesion	O	O	B-anat	I-anat	I-anat	I-anat	I-anat	L-anat	O

nodule

monocytes

Classic named entity recognition

➔ If we have **overlapping**/**nested** entities, the classic flat tag schemes do not work anymore

La patiente a un petit nodule dans le quadrant sup. ext. du sein droit .

(= The patient has a small nodule in the upper outer quadrant of the right breast.)

La	patiente	a	un	petit	nodule	dans	le	quadrant	sup.	ext.	du	sein	droit	.
⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

Multiple options

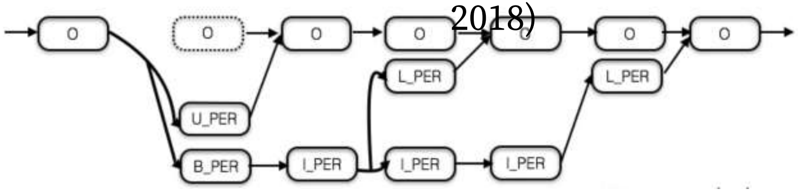
Several options to predict overlapping entities

➔ adapt the tagging scheme or but there are limitations

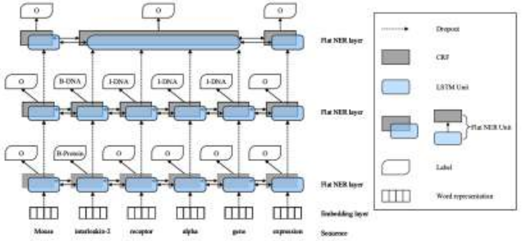
➔ predict non overlapping entities layer by layer (small first, then larger …)
  - can we improve this approach by preventing layer specialization ?

➔ do away with token classification and classify each possible span instead
  - can we keep a token classification step ?

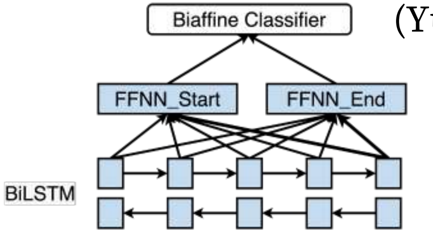
(Katiyar et al., 2018)



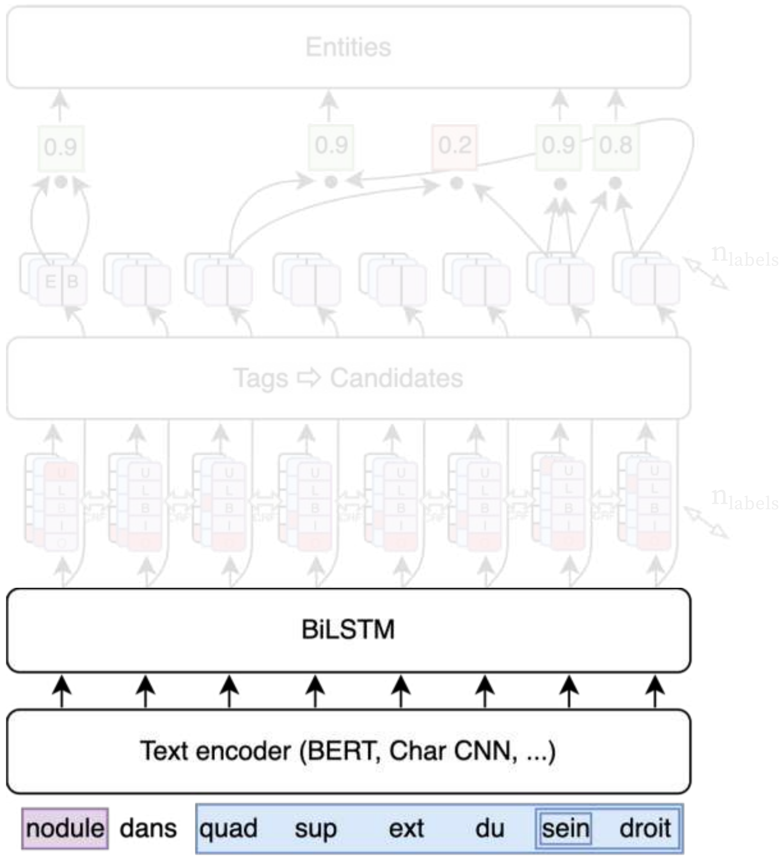
(Ju et al., 2018)



(Yu et al., 2020)

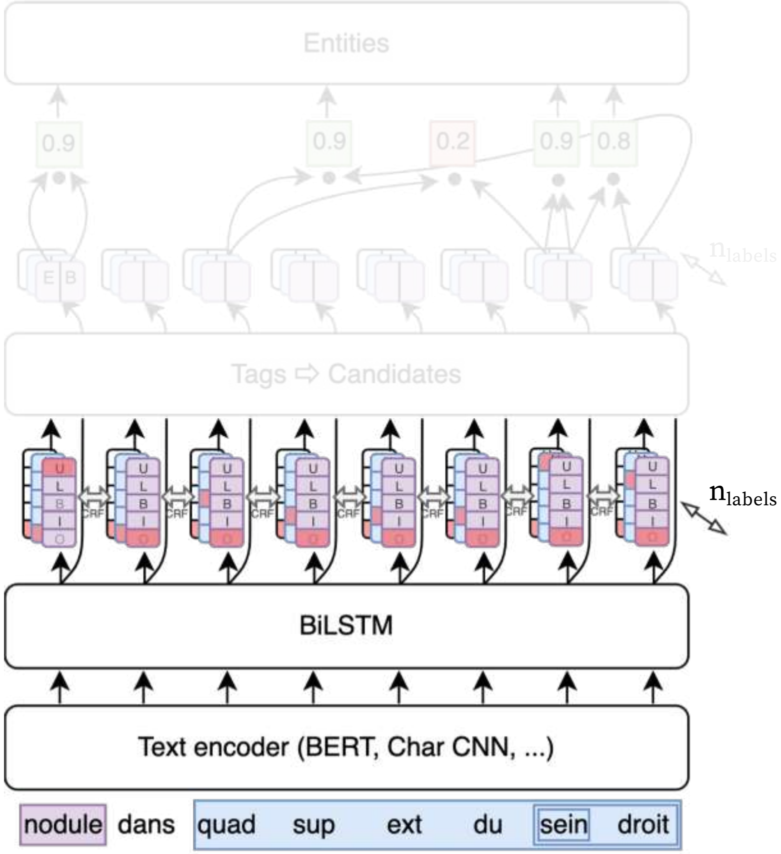


Method 1: tag and filter model



- ➔ Encode the text as embeddings
- ➔ Predict **multiple** labels per word
- ➔ Convert into candidate entities
- ➔ Filter these candidates, but keep at least one candidate for each non empty word

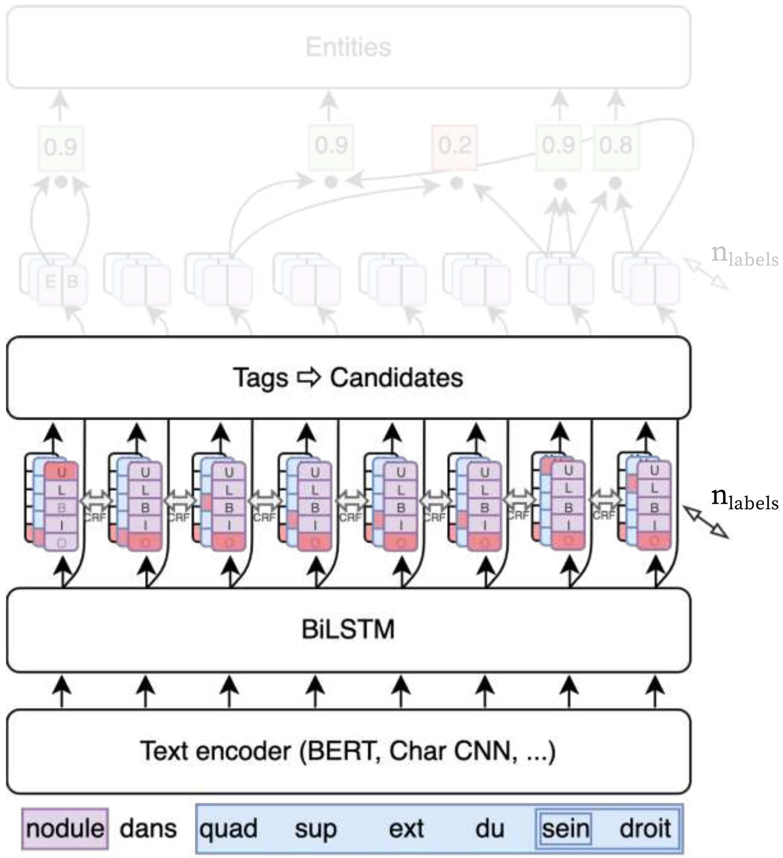
Method 1: tag and filter model



- ➔ Encode the text as embeddings
- ➔ Predict **multiple** labels per word
- ➔ Convert into candidate entities
- ➔ Filter these candidates, but keep at least one candidate for each non empty word

nodule	dans	quad.	sup.	ext.	du	sein	droit
⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩
O	O	B-anat	I-anat	I-anat	I-anat	U-anat	L-anat
U-lesion	O	O	O	O	O	O	O

Method 1: tag and filter model



- ➔ Encode the text as embeddings
- ➔ Predict **multiple** labels per word
- ➔ Convert into candidate entities
- ➔ Filter these candidates, but keep at least one candidate for each non empty word

monocytes ⇨

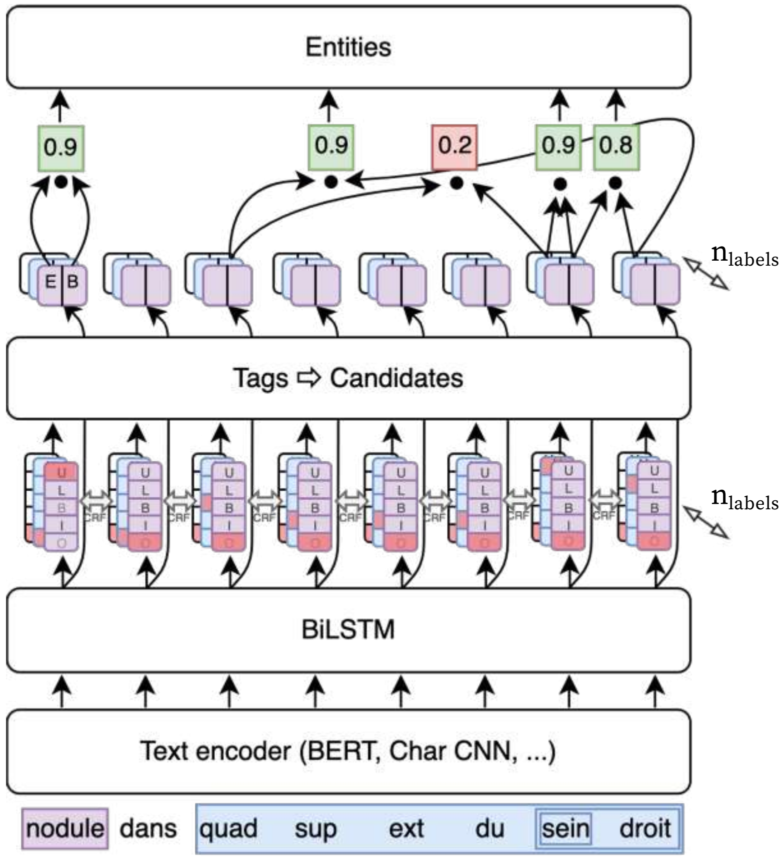
⇨

quadrant sup ext du sein droit ⇨

quad sup ext du sein ⇨

sein ⇨

Method 1: tag and filter model



- ➔ Encode the text as embeddings
- ➔ Predict **multiple** labels per word
- ➔ Convert into candidate entities
- ➔ Filter these candidates, but keep at least one candidate for each non empty word

monocytes ⇨

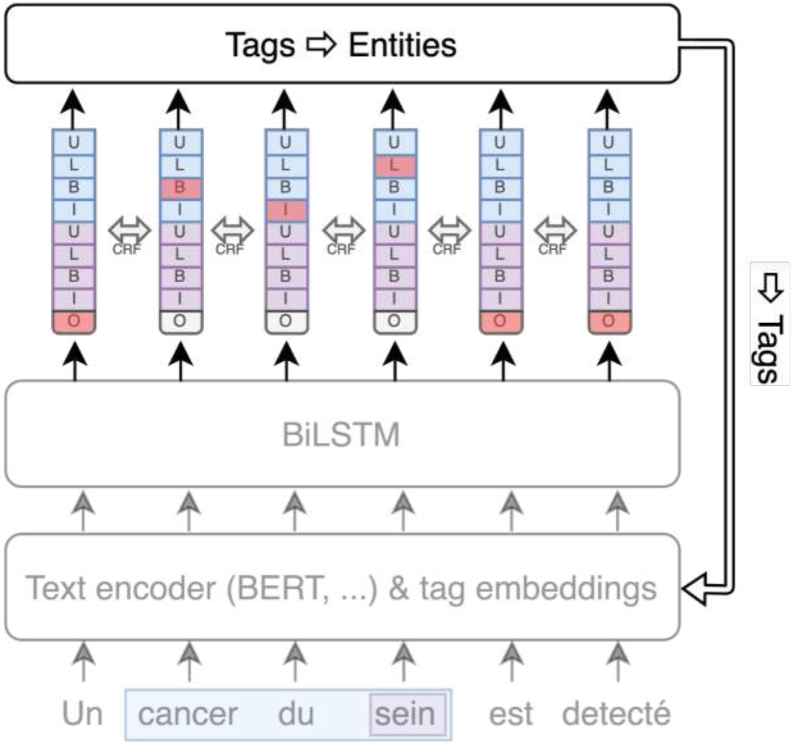
✓ ⇨

quadrant sup ext du sein droit ⇨

✓ ⇨

quad sup ext du sein ⇨

Method 2: autoregressive model



➔ Iteratively predict non overlapping entities

➔ Feed back predictions to the model to prevent repetition and improve next predictions

➔ Train the model with custom strategies

Example (train to predict large entities first)

Init: no entities

Step 1: predict cancer du sein

Step 2: predict sein

Step 3: predict Ø ➔ we stop

Experiments: datasets

	DEFT 3.1		DEFT 3.2		GENIA			CONLL EN 2003		
	train	test	train	test	train	val	test	train	val	test
Language	FR		FR		EN			EN		
Domain	Clinical		Clinical		Biomedical			General		
# docs	100	67	100	67	1599	190	213	946	216	231
# entities	5677		2167	1445	46185	4379	5515	23499	5942	5648
avg length	1.94	2.03	4.55	4.74	1.90	2.11	2.05	1.45	1.45	1.44
# unique labels	8	8	2	2	5	5	5	4	4	4
# unique texts	3449	2179	1878	1320	15441	2141	2681	8082	2809	2637
# nestings	475	422	14	4	4524	436	658	0	0	0
# same label nestings	8	2	2	1	2430	234	331	0	0	0
# crossing overlaps	1	0	0	0	0	0	0	0	0	0
# same label crossing	0	0	0	0	0	0	0	0	0	0
# superpositions	0	1	0	0	43	12	9	0	0	0

- ➔ French and English
- ➔ General and medical
- ➔ Small and large
- ➔ With more or less overlap

Experiments: datasets

	DEFT 3.1		DEFT 3.2		GENIA			CONLL EN 2003		
	train	test	train	test	train	val	test	train	val	test
Language	FR		FR		EN			EN		
Domain	Clinical		Clinical		Biomedical			General		
# docs	100	67	100	67	1599	190	213	946	216	231
# entities	5677		2167	1445	46185	4379	5515	23499	5942	5648
avg length	1.94	2.03	4.55	4.74	1.90	2.11	2.05	1.45	1.45	1.44
# unique labels	8	8	2	2	5	5	5	4	4	4
# unique texts	3449	2179	1878	1320	15441	2141	2681	8082	2809	2637
# nestings	475	422	14	4	4524	436	658	0	0	0
# same label nestings	8	2	2	1	2430	234	331	0	0	0
# crossing overlaps	1	0	0	0	0	0	0	0	0	0
# same label crossing	0	0	0	0	0	0	0	0	0	0
# superpositions	0	1	0	0	43	12	9	0	0	0

➔ French and English

➔ General and medical

➔ Small and large

➔ With more or less overlap

Experiments: datasets

	DEFT 3.1		DEFT 3.2		GENIA			CONLL EN 2003		
	train	test	train	test	train	val	test	train	val	test
Language	FR		FR		EN			EN		
Domain	Clinical		Clinical		Biomedical			General		
# docs	100	67	100	67	1599	190	213	946	216	231
# entities	5677		2167	1445	46185	4379	5515	23499	5942	5648
avg length	1.94	2.03	4.55	4.74	1.90	2.11	2.05	1.45	1.45	1.44
# unique labels	8	8	2	2	5	5	5	4	4	4
# unique texts	3449	2179	1878	1320	15441	2141	2681	8082	2809	2637
# nestings	475	422	14	4	4524	436	658	0	0	0
# same label nestings	8	2	2	1	2430	234	331	0	0	0
# crossing overlaps	1	0	0	0	0	0	0	0	0	0
# same label crossing	0	0	0	0	0	0	0	0	0	0
# superpositions	0	1	0	0	43	12	9	0	0	0

- ➔ French and English
- ➔ General and medical
- ➔ Small and large
- ➔ With more or less overlap

Experiments: datasets

	DEFT 3.1		DEFT 3.2		GENIA			CONLL EN 2003		
	train	test	train	test	train	val	test	train	val	test
Language	FR		FR		EN			EN		
Domain	Clinical		Clinical		Biomedical			General		
# docs	100	67	100	67	1599	190	213	946	216	231
# entities	5677		2167	1445	46185	4379	5515	23499	5942	5648
avg length	1.94	2.03	4.55	4.74	1.90	2.11	2.05	1.45	1.45	1.44
# unique labels	8	8	2	2	5	5	5	4	4	4
# unique texts	3449	2179	1878	1320	15441	2141	2681	8082	2809	2637
# nestings	475	422	14	4	4524	436	658	0	0	0
# same label nestings	8	2	2	1	2430	234	331	0	0	0
# crossing overlaps	1	0	0	0	0	0	0	0	0	0
# same label crossing	0	0	0	0	0	0	0	0	0	0
# superpositions	0	1	0	0	43	12	9	0	0	0

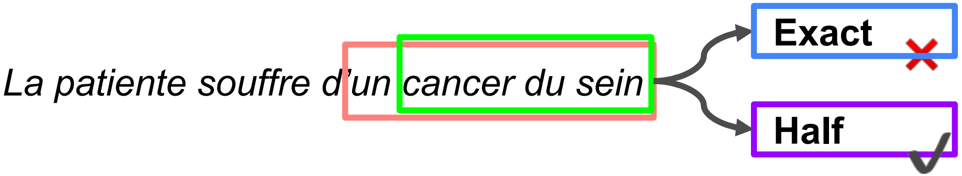
- ➔ French and english
- ➔ General and medical
- ➔ ~~Small and large~~
- ➔ With more or less overlap

Experiments: general results

DEFT (F1)	Exact	Half
Copara et al. [2020]	70.7	
Copara et al. [2020] (× 3)	72.6	
BERT + softmax	50.4	60.5
Autoreg short→large	74.1	84.5
BiTag w/o finetuning	73.9	83.6
Biaffine only	73.5	82.1
BiTag	74.3	84.3
Autoreg short→large (× 3)	75.4	85.2
BiTag (× 3)	75.3	85.4

GENIA (F1)	Exact	Half
Lin et al. [2019]	74.8	
Shibuya and Hovy [2020]	75.5	
Luan et al. [2019]	76.2	
Straková et al. [2019]*	78.3	
Wang et al. [2020]*	79.3	
BERT + softmax	73.8	81.7
Autoreg large→short	78.3	84.3
BiTag w/o fine-tuning	78.1	83.4
Biaffine-only	78.5	83.8
BiTag	78.4	84.3
Autoreg large→short (× 3)	79.0	85.1
BiTag (× 3)	79.1	85.1

CoNLL (F1)	Exact	Half
Lample et al. [2016]	90.9	
Strubell et al. [2017]	90.7	
Devlin et al. [2019]	92.8	
Straková et al. [2019]	93.4	
Yu et al. [2020]	93.5	
BERT + softmax	91.1	92.8
Autoreg	93.0	94.2
BiTag w/o finetuning	92.6	94.1
Biaffine-only	92.8	94.0
BiTag	93.1	94.3
Autoreg (× 3)	93.6	94.5
BiTag (× 3)	93.4	94.7



Experiments: general results

DEFT (F1)	Exact	Half
Copara et al. [2020]	70.7	
Copara et al. [2020] (× 3)	72.6	
BERT + softmax	50.4	60.5
Autoreg short→large	74.1	84.5
BiTag w/o finetuning	73.9	83.6
Biaffine only	73.5	82.1
BiTag	74.3	84.3
Autoreg short→large (× 3)	75.4	85.2
BiTag (× 3)	75.3	85.4

GENIA (F1)	Exact	Half
Lin et al. [2019]	74.8	
Shibuya and Hovy [2020]	75.5	
Luan et al. [2019]	76.2	
Straková et al. [2019]*	78.3	
Wang et al. [2020]*	79.3	
BERT + softmax	73.8	81.7
Autoreg large→short	78.3	84.3
BiTag w/o fine-tuning	78.1	83.4
Biaffine-only	78.5	83.8
BiTag	78.4	84.3
Autoreg large→short (× 3)	79.0	85.1
BiTag (× 3)	79.1	85.1

CoNLL (F1)	Exact	Half
Lample et al. [2016]	90.9	
Strubell et al. [2017]	90.7	
Devlin et al. [2019]	92.8	
Straková et al. [2019]	93.4	
Yu et al. [2020]	93.5	
BERT + softmax	91.1	92.8
Autoreg	93.0	94.2
BiTag w/o finetuning	92.6	94.1
Biaffine-only	92.8	94.0
BiTag	93.1	94.3
Autoreg (× 3)	93.6	94.5
BiTag (× 3)	93.4	94.7

➔ The classic token classification model fails on nested datasets

➔ No noticeable difference between the two proposed models

➔ Discrepancy between **Exact** and **Half** and token classification helps **Half F1**

➔ Ensembling improves the performance

Experiments: general results

DEFT (F1)	Exact	Half
Copara et al. [2020]	70.7	
Copara et al. [2020] (× 3)	72.6	
BERT + softmax	50.4	60.5
Autoreg short→large	74.1	84.5
BiTag w/o finetuning	73.9	83.6
Biaffine only	73.5	82.1
BiTag	74.3	84.3
Autoreg short→large (× 3)	75.4	85.2
BiTag (× 3)	75.3	85.4

GENIA (F1)	Exact	Half
Lin et al. [2019]	74.8	
Shibuya and Hovy [2020]	75.5	
Luan et al. [2019]	76.2	
Straková et al. [2019]*	78.3	
Wang et al. [2020]*	79.3	
BERT + softmax	73.8	81.7
Autoreg large→short	78.3	84.3
BiTag w/o fine-tuning	78.1	83.4
Biaffine-only	78.5	83.8
BiTag	78.4	84.3
Autoreg large→short (× 3)	79.0	85.1
BiTag (× 3)	79.1	85.1

CoNLL (F1)	Exact	Half
Lample et al. [2016]	90.9	
Strubell et al. [2017]	90.7	
Devlin et al. [2019]	92.8	
Straková et al. [2019]	93.4	
Yu et al. [2020]	93.5	
BERT + softmax	91.1	92.8
Autoreg	93.0	94.2
BiTag w/o finetuning	92.6	94.1
Biaffine-only	92.8	94.0
BiTag	93.1	94.3
Autoreg (× 3)	93.6	94.5
BiTag (× 3)	93.4	94.7

$\simeq$

- ➔ The classic token classification model fails on nested datasets
- ➔ No noticeable difference between the two proposed models
- ➔ Discrepancy between **Exact** and **Half** and token classification helps **Half F1**
- ➔ Ensembling improves the performance

Nested named entity recognition

Experiments: general results

DEFT (F1)	Exact	Half
Copara et al. [2020]	70.7	
Copara et al. [2020] (× 3)	72.6	
BERT + softmax	50.4	60.5
Autoreg short→large	74.1	84.5
BiTag w/o finetuning	73.9	83.6
Biaffine only	73.5	82.1
BiTag	74.3	84.3
Autoreg short→large (× 3)	75.4	85.2
BiTag (× 3)	75.3	85.4

+0.8 +2.2

GENIA (F1)	Exact	Half
Lin et al. [2019]	74.8	
Shibuya and Hovy [2020]	75.5	
Luan et al. [2019]	76.2	
Straková et al. [2019]*	78.3	
Wang et al. [2020]*	79.3	
BERT + softmax	73.8	81.7
Autoreg large→short	78.3	84.3
BiTag w/o fine-tuning	78.1	83.4
Biaffine-only	78.5	83.8
BiTag	78.4	84.3
Autoreg large→short (× 3)	79.0	85.1
BiTag (× 3)	79.1	85.1

-0.1 +0.5

CoNLL (F1)	Exact	Half
Lample et al. [2016]	90.9	
Strubell et al. [2017]	90.7	
Devlin et al. [2019]	92.8	
Straková et al. [2019]	93.4	
Yu et al. [2020]	93.5	
BERT + softmax	91.1	92.8
Autoreg	93.0	94.2
BiTag w/o finetuning	92.6	94.1
Biaffine-only	92.8	94.0
BiTag	93.1	94.3
Autoreg (× 3)	93.6	94.5
BiTag (× 3)	93.4	94.7

- ➔ The classic token classification model fails on nested datasets
- ➔ No noticeable difference between the two proposed models
- ➔ Discrepancy between **Exact** and **Half** and token classification helps **Half F1**
- ➔ Ensembling improves the performance

Experiments: general results

DEFT (F1)	Exact	Half
Copara et al. [2020]	70.7	
Copara et al. [2020] (× 3)	72.6	
BERT + softmax	50.4	60.5
Autoreg short→large	74.1	84.5
BiTag w/o finetuning	73.9	83.6
Biaffine only	73.5	82.1
BiTag	74.3	84.3
Autoreg short→large (× 3)	75.4	85.2
BiTag (× 3)	75.3	85.4

GENIA (F1)	Exact	Half
Lin et al. [2019]	74.8	
Shibuya and Hovy [2020]	75.5	
Luan et al. [2019]	76.2	
Straková et al. [2019]*	78.3	
Wang et al. [2020]*	79.3	
BERT + softmax	73.8	81.7
Autoreg large→short	78.3	84.3
BiTag w/o fine-tuning	78.1	83.4
Biaffine-only	78.5	83.8
BiTag	78.4	84.3
Autoreg large→short (× 3)	79.0	85.1
BiTag (× 3)	79.1	85.1

CoNLL (F1)	Exact	Half
Lample et al. [2016]	90.9	
Strubell et al. [2017]	90.7	
Devlin et al. [2019]	92.8	
Straková et al. [2019]	93.4	
Yu et al. [2020]	93.5	
BERT + softmax	91.1	92.8
Autoreg	93.0	94.2
BiTag w/o finetuning	92.6	94.1
Biaffine-only	92.8	94.0
BiTag	93.1	94.3
Autoreg (× 3)	93.6	94.5
BiTag (× 3)	93.4	94.7

➔ The classic token classification model fails on nested datasets

➔ No noticeable difference between the two proposed models

➔ Discrepancy between **Exact** and **Half** and token classification helps **Half F1**

➔ Ensembling improves the performance

Experiments: ablations

We document other findings by ablating parts of our models:

- ➔ finetuning the encoder weights can be beneficial
- ➔ adding surrounding context to BERT embeddings improves the performance
- ➔ optimal autoregressive order varies with the dataset
- ➔ BIOUL encoding scheme is best both for decoding and encoding entities

	DEFT		GENIA	
	Exact	Half	Exact	Half
base	71.4	80.9	78.9	84.5
– Tagging	71.2 (–0.2)	79.2 (–1.7)	78.8 (–0.1)	83.5 (–1.0)
– Doc context	70.6 (–0.8)	80.2 (–0.7)	78.6 (–0.3)	85.0 (–0.2)
– Char CNN	71.0 (–0.4)	80.2 (–0.7)	78.8 (–0.1)	84.4 (–0.1)
– FastText	71.8 (+0.4)	81.1 (+0.2)	78.8 (–0.1)	84.4 (–0.1)
+ Finetuning	73.3 (+1.9)	82.4 (+1.5)	78.9 (+0.0)	84.5 (+0.0)

	DEFT		GENIA	
	Exact	Half	Exact	Half
large → short	70.5	79.7	**79.5**	**85.2**
greedy	71.1	80.3	79.2	**85.2**
short → large	**71.6**	**80.6**	78.7	85.0

DEFT	BIO encoding	BIOUL encoding
BIO decoding	70.1	71.3
BIOUL decoding	70.5	**71.6**

Experiments: ablations

We document other findings by ablating parts of our models:

➔ finetuning the encoder weights can be beneficial

➔ adding surrounding context to BERT embeddings improves the performance

➔ optimal autoregressive order varies with the dataset

➔ BIOUL encoding scheme is best both for decoding and encoding entities

	DEFT		GENIA	
	Exact	Half	Exact	Half
base	71.4	80.9	78.9	84.5
– Tagging	71.2 (–0.2)	79.2 (–1.7)	78.8 (–0.1)	83.5 (–1.0)
– Doc context	70.6 (–0.8)	80.2 (–0.7)	78.6 (–0.3)	85.0 (–0.2)
– Char CNN	71.0 (–0.4)	80.2 (–0.7)	78.8 (–0.1)	84.4 (–0.1)
– FastText	71.8 (+0.4)	81.1 (+0.2)	78.8 (–0.1)	84.4 (–0.1)
+ Finetuning	73.3 (+1.9)	82.4 (+1.5)	78.9 (+0.0)	84.5 (+0.0)

	DEFT		GENIA	
	Exact	Half	Exact	Half
large → short	70.5	79.7	**79.5**	**85.2**
greedy	71.1	80.3	79.2	**85.2**
short → large	**71.6**	**80.6**	78.7	85.0

DEFT	BIO encoding	BIOUL encoding
BIO decoding	70.1	71.3
BIOUL decoding	70.5	**71.6**

Experiments: ablations

We document other findings by ablating parts of our models:

- ➔ finetuning the encoder weights can be beneficial
- ➔ adding surrounding context to BERT embeddings improves the performance
- ➔ optimal autoregressive order varies with the dataset
- ➔ BIOUL encoding scheme is best both for decoding and encoding entities

	DEFT		GENIA	
	Exact	Half	Exact	Half
base	71.4	80.9	78.9	84.5
– Tagging	71.2 (–0.2)	79.2 (–1.7)	78.8 (–0.1)	83.5 (–1.0)
– Doc context	70.6 (–0.8)	80.2 (–0.7)	78.6 (–0.3)	85.0 (–0.2)
– Char CNN	71.0 (–0.4)	80.2 (–0.7)	78.8 (–0.1)	84.4 (–0.1)
– FastText	71.8 (+0.4)	81.1 (+0.2)	78.8 (–0.1)	84.4 (–0.1)
+ Finetuning	73.3 (+1.9)	82.4 (+1.5)	78.9 (+0.0)	84.5 (+0.0)

	DEFT		GENIA	
	Exact	Half	Exact	Half
large → short	70.5	79.7	**79.5**	**85.2**
greedy	71.1	80.3	79.2	**85.2**
short → large	**71.6**	**80.6**	78.7	85.0

DEFT	BIO encoding	BIOUL encoding
BIO decoding	70.1	71.3
BIOUL decoding	70.5	**71.6**

Experiments: ablations

We document other findings by ablating parts of our models:

- ➔ finetuning the encoder weights can be beneficial
- ➔ adding surrounding context to BERT embeddings improves the performance
- ➔ optimal autoregressive order varies with the dataset
- ➔ BIOUL encoding scheme is best both for decoding and **encoding** entities

	DEFT		GENIA	
	Exact	Half	Exact	Half
base	71.4	80.9	78.9	84.5
– Tagging	71.2 (+0.2)	79.2 (–1.7)	78.8 (–0.1)	83.5 (–1.0)
– Doc context	70.6 (–0.8)	80.2 (–0.7)	78.6 (–0.3)	85.0 (–0.2)
– Char CNN	71.0 (–0.4)	80.2 (–0.7)	78.8 (–0.1)	84.4 (–0.1)
– FastText	71.8 (+0.4)	81.1 (+0.2)	78.8 (–0.1)	84.4 (–0.1)
+ Finetuning	73.3 (+1.9)	82.4 (+1.5)	78.9 (+0.0)	84.5 (+0.0)

	DEFT		GENIA	
	Exact	Half	Exact	Half
large → short	70.5	79.7	**79.5**	**85.2**
greedy	71.1	80.3	79.2	**85.2**
short → large	**71.6**	**80.6**	78.7	85.0

DEFT	BIO encoding	BIOUL encoding
BIO decoding	70.1	71.3
BIOUL decoding	70.5	**71.6**

Key contributions & findings

- ➔ Two methods for overlapping NER
- ➔ Features matter: finetune BERT and add surrounding context
- ➔ Tag classification helps, especially w.r.t. relaxed match performance
- ➔ Optimal autoregressive order can vary depending on the dataset
- ➔ Exact match metric should be completed by a relaxed metric

Multilingual medical named entity normalization

A retrieval and translation problem

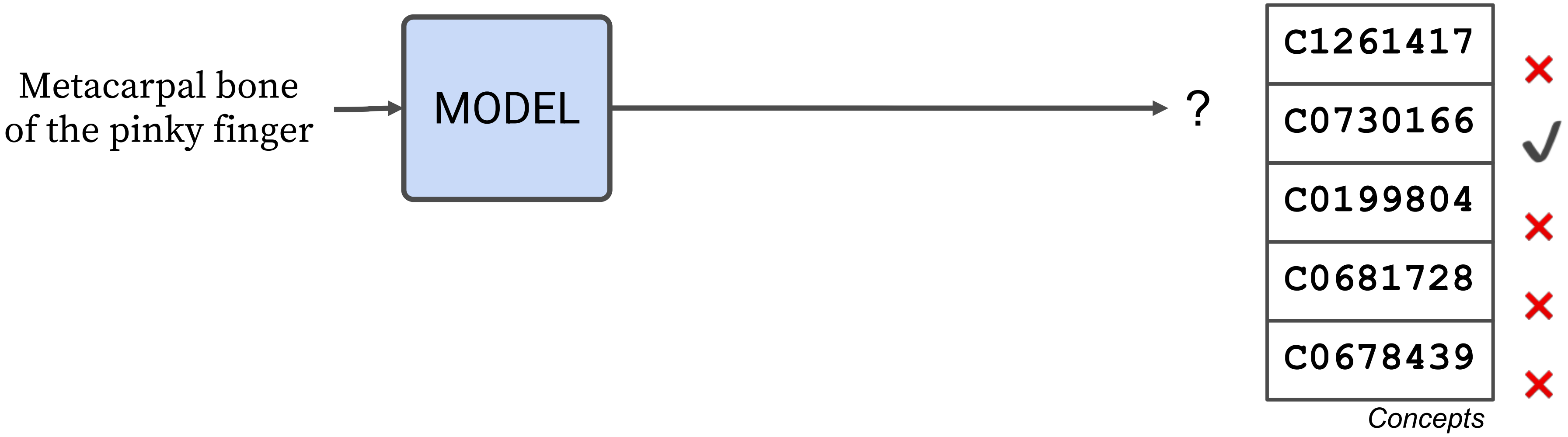
➔ A **terminology**, such as the **UMLS**, contains **concepts** and at least one example/**synonym** for each concept

Synonyms	***Concepts***
5th metacarpal bone	C0730166
bcg vaccination	C0199804
kidney transplant	C1261317
fifth metacarpal bone	C0730166
bone of the 5th me...	C0730166
café robusta	C0678439
attention	C0004268

...

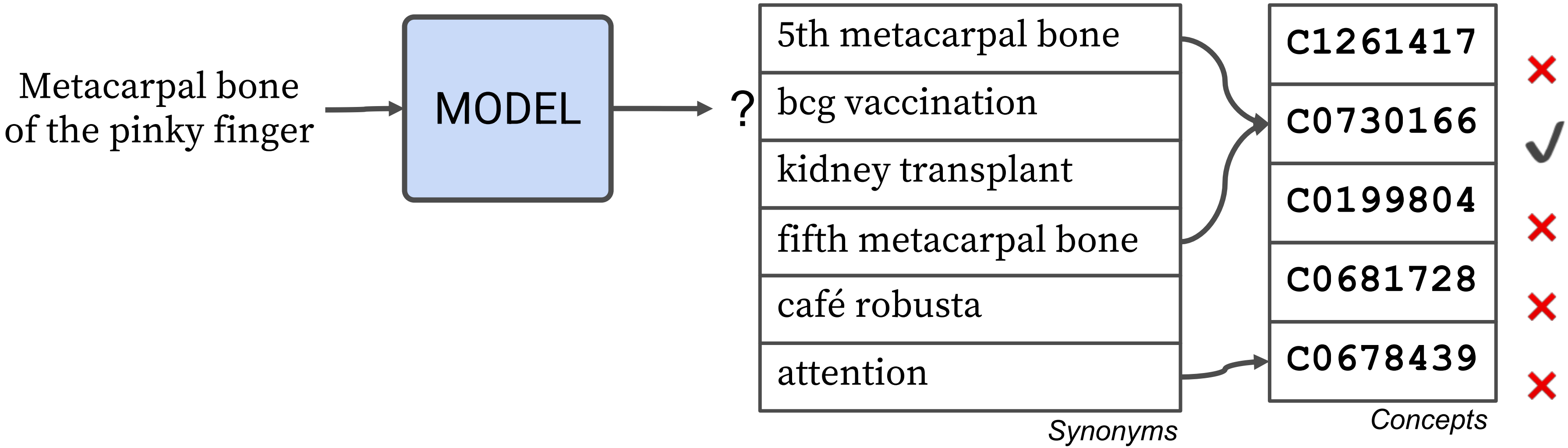
A retrieval and translation problem

- Given an **extracted named entity**, map it to the correct **concept** in a terminology
- Some methods directly classify the named entity
- But only English works and medium-sized terminologies < 160 000 concepts



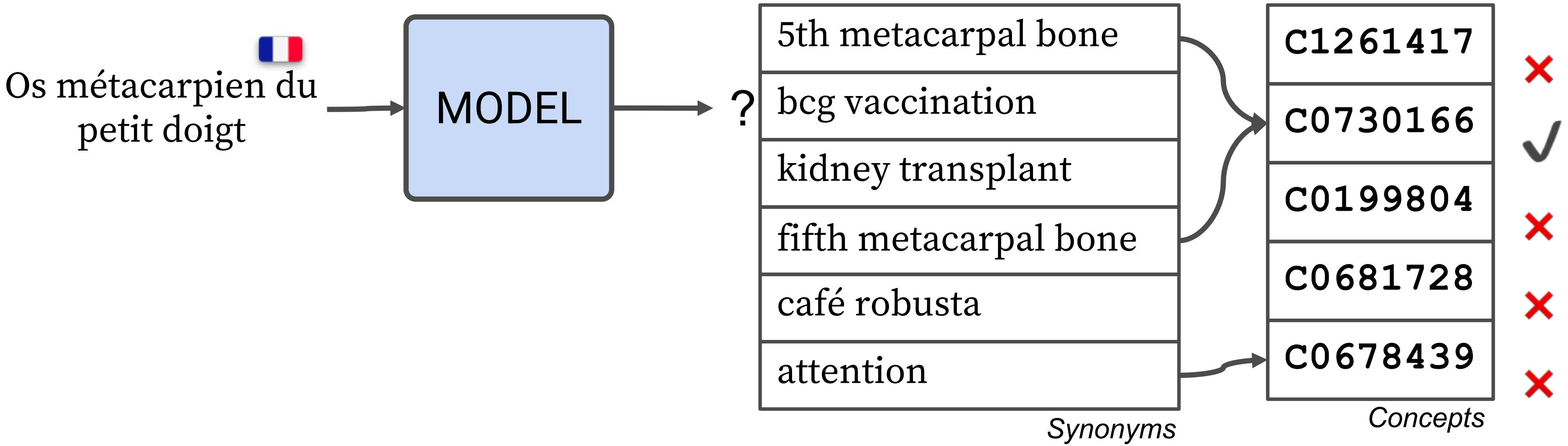
A retrieval and translation problem

- ➔ Most methods search the closest synonym and lookup its concept
- ➔ But this means **larger/slower** models since each synonym needs to be embedded



A retrieval and translation problem

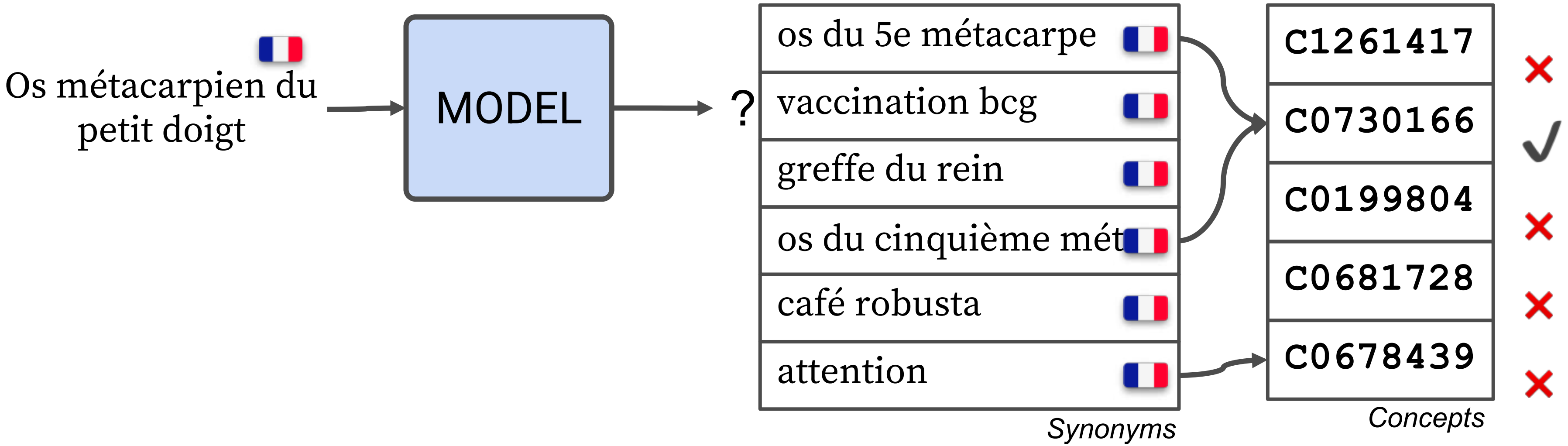
➔ What if the source and target **languages differ** ?



A retrieval and translation problem

➔ What if the source and target **languages differ** ?

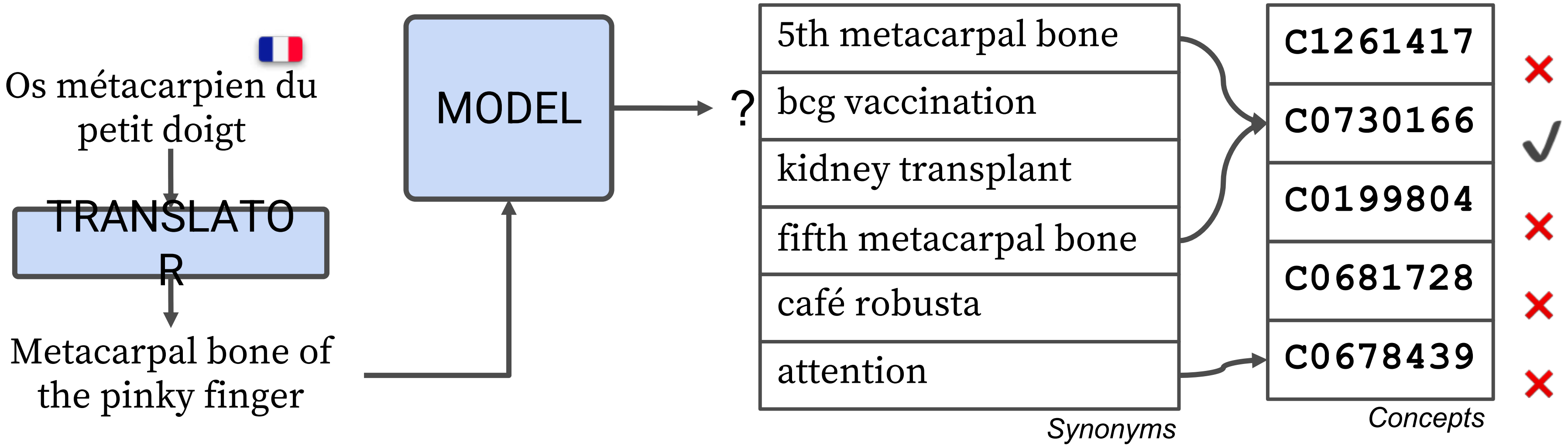
➔ Existing literature relies synonym lookup with manual or machine **translation of terminologies**...



A retrieval and translation problem

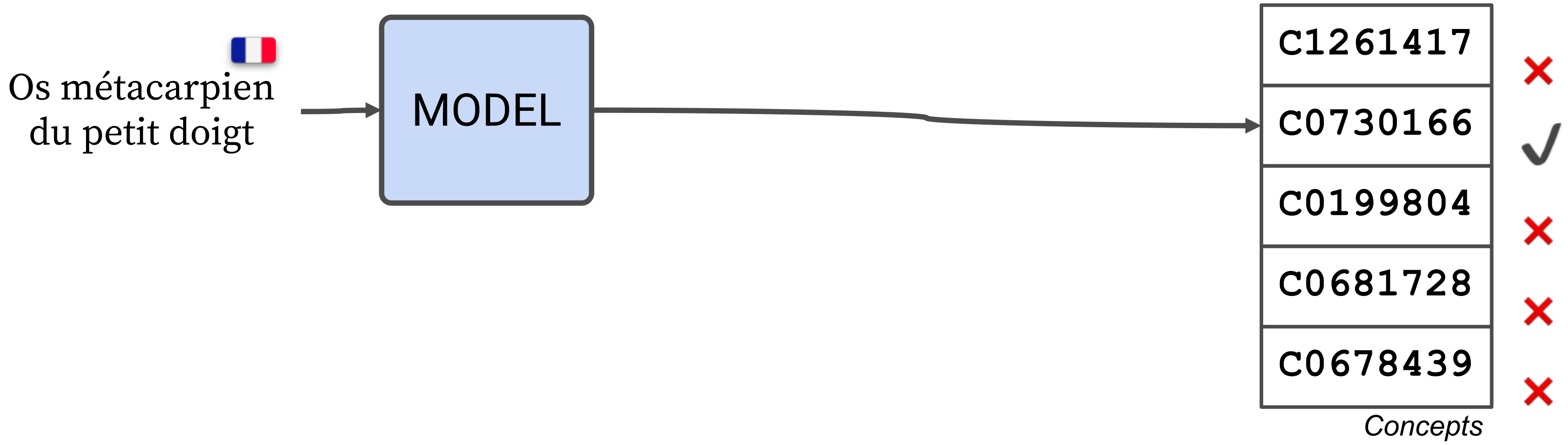
➔ … or the **translation of named entities** before normalizing them *(Roller et al., 2018)*

➔ However, doing this can be a **source of error** and makes models **more complex** or dependent on **external** services



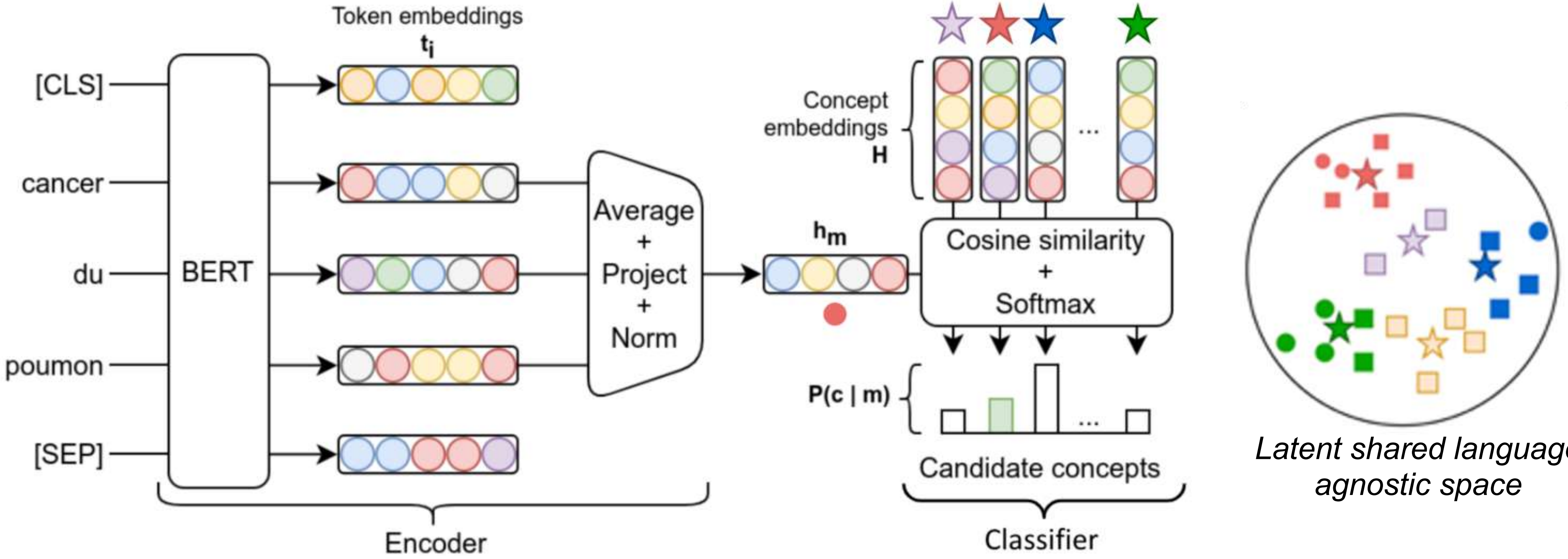
A retrieval and translation problem

➔ Could we **skip** all these steps and still normalize named entities in **non-English** languages against **large** terminologies ?



Architecture of our classifier

➔ Embed synonyms and concepts in a shared and language-agnostic space



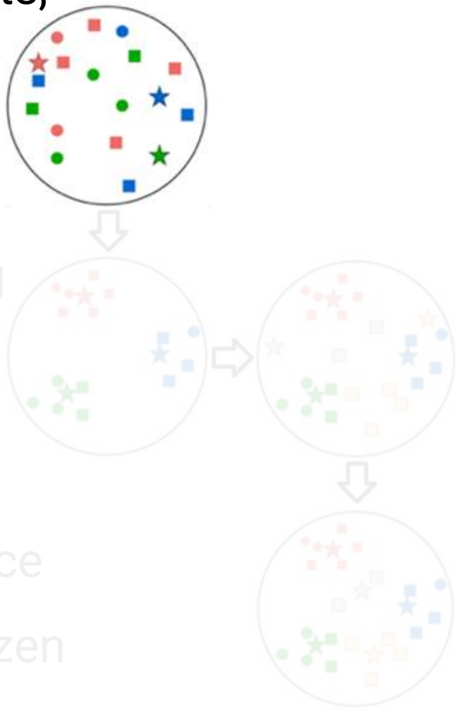
Latent shared language agnostic space

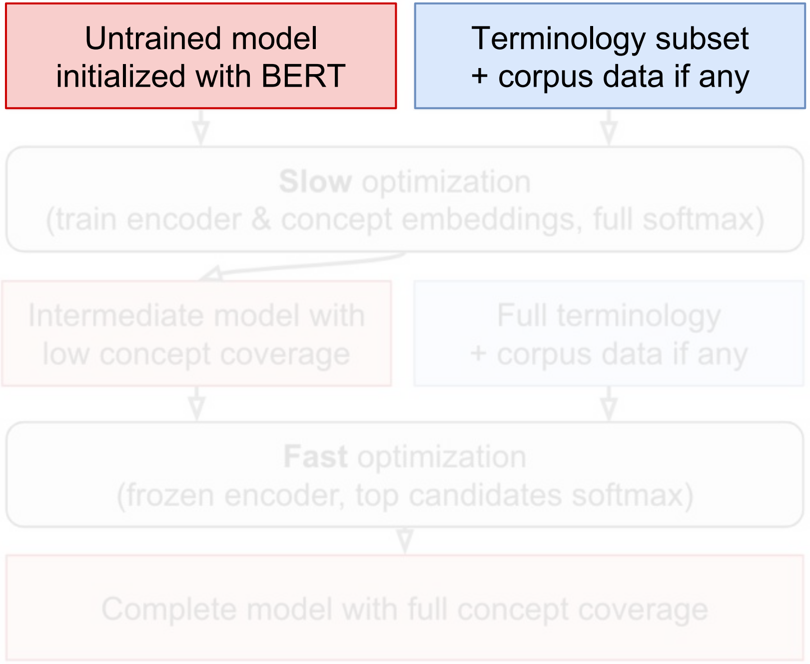
Two steps training

➔ Too many concepts, so we train on a subset first

➔ Then, add missing concepts

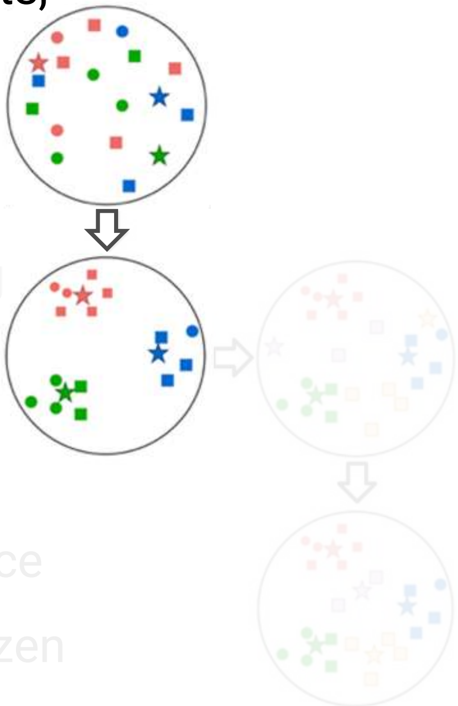
➔ Benefits optimizations, since the encoder is frozen

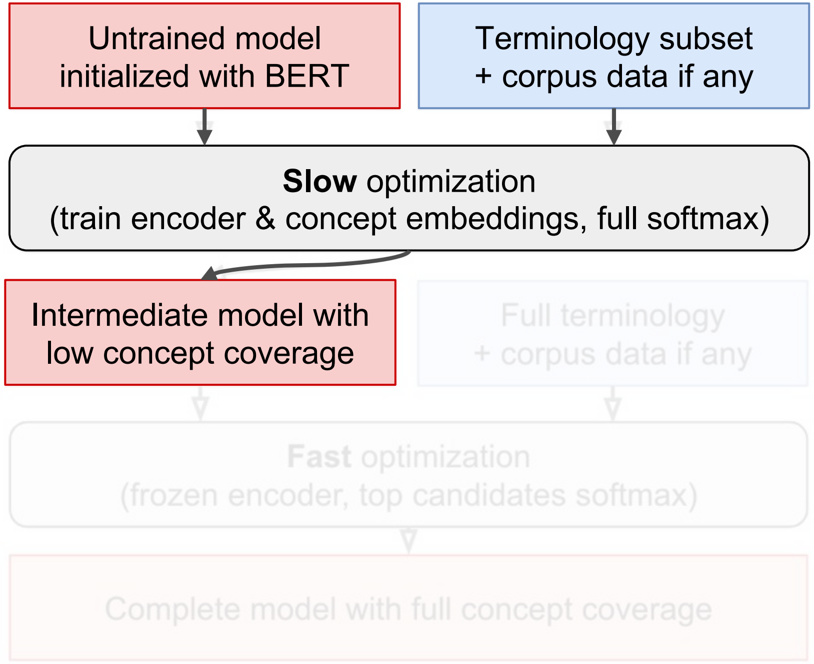




Two steps training

- ➔ Too many concepts, so we train on a subset first
- ➔ Then, add missing concepts
- ➔ Benefits optimizations, since the encoder is frozen



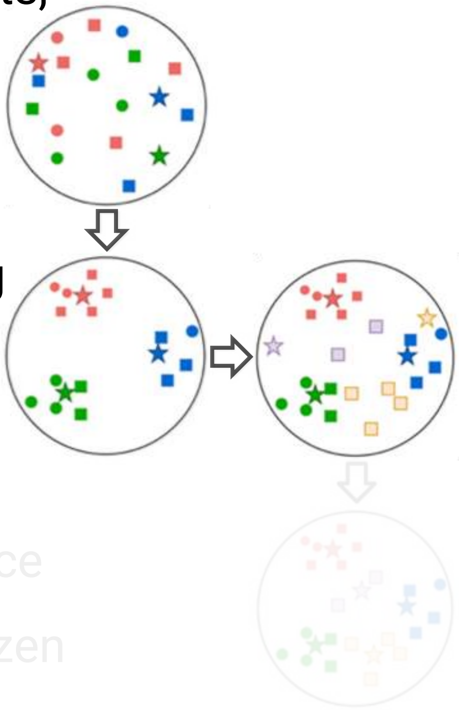


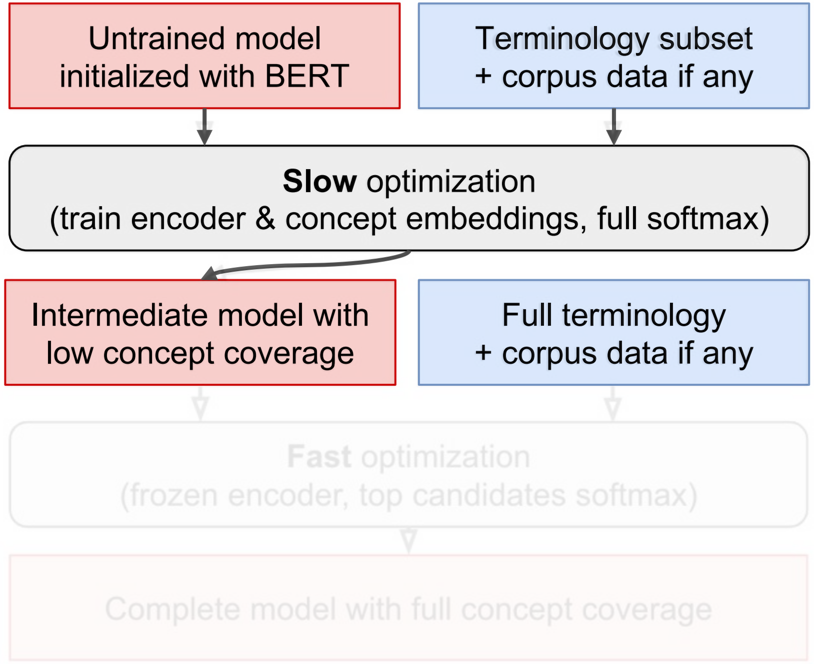
Two steps training

➔ Too many concepts, so we train on a subset first

➔ Then, add missing concepts

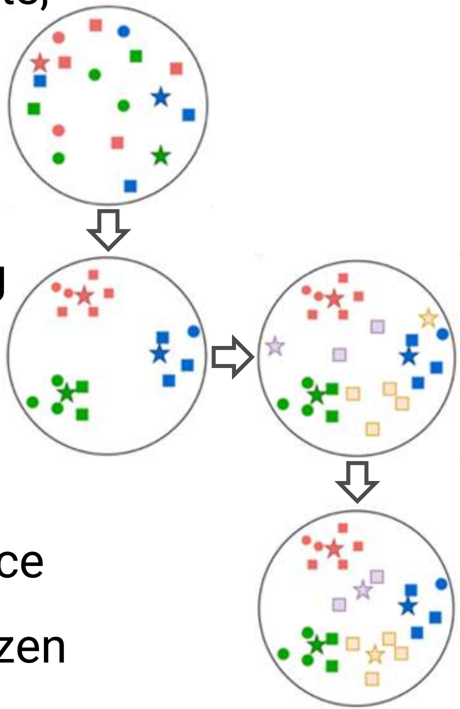
➔ Benefits optimizations, since the encoder is frozen

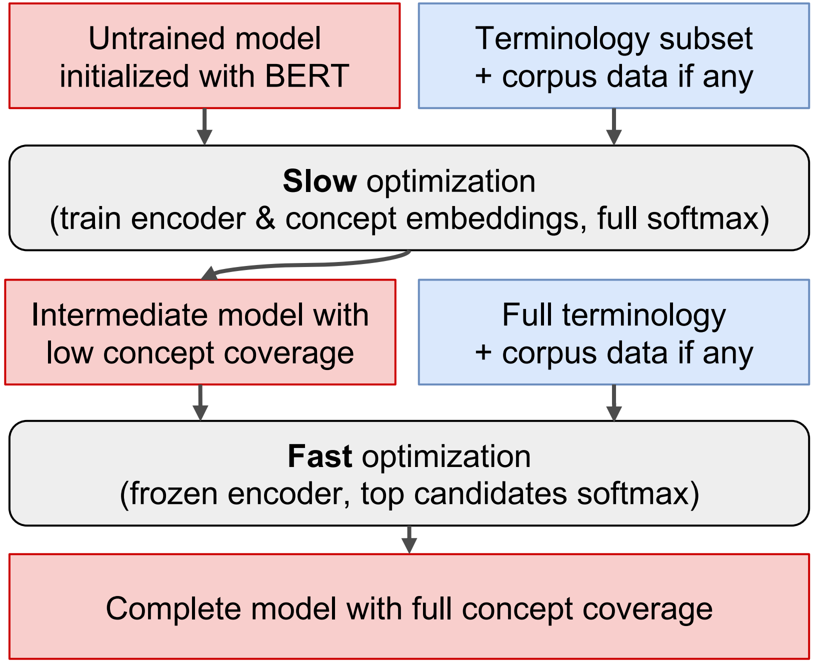




Two steps training

- ➔ Too many concepts, so we train on a subset first
- ➔ Then, add missing concepts
- ➔ Benefits optimizations, since the encoder is frozen





Experiments: data

➔ **Quaero dataset** (2015 & 2016 versions)

Language: *French*

Terminology: *Filtered UMLS = 766 548 concepts*

Coverage: *~70% of concepts in French*

Annotated training samples: *5695*

➔ **Mantra dataset:**

Languages: *English, French, Spanish, Dutch and German*

Terminology: *Mantra terminology = 591 918 concepts*

Coverage: *~65% in French/Dutch/German, 93% in Spanish, 100% for English*

Annotated training samples: *0*

Experiments: general results

Quaero (F1)

		Quaero 2015		Quaero 2016	
		MEDLINE	EMEA	MEDLINE	EMEA
Others	[Afzal et al., 2015]	67.1	**87.2**	—	—
	[Cabot et al., 2016]	—	—	55.2	52.4
	[Roller et al., 2018]	73.6	83.5	71.3	73.4
Our model	no corpus annotations	73.7	76.5	75.4	72.7
	with corpus annotations	**79.0**	85.1	**79.0**	**74.3**

Mantra Medline (F1)

	English	Spanish	French	Dutch	German
[Roller et al., 2018]	—	68.7	68.6	64.8	67.9
Our model	81.7	**74.5**	**71.5**	**70.0**	**76.0**

➔ compares favorably to the state of the art on both datasets

➔ even good results with no corpus training annotations for Quaero

Experiments: general results

Quaero (F1)

		Quaero 2015		Quaero 2016	
		MEDLINE	EMEA	MEDLINE	EMEA
Others	[Afzal et al., 2015]	67.1	**87.2**	—	—
	[Cabot et al., 2016]	—	—	55.2	52.4
	[Roller et al., 2018]	73.6	83.5	71.3	73.4
Our model	no corpus annotations	73.7	76.5	75.4	72.7
	with corpus annotations	**79.0**	85.1	**79.0**	**74.3**

Mantra Medline (F1)

	English	Spanish	French	Dutch	German
[Roller et al., 2018]	—	68.7	68.6	64.8	67.9
Our model	81.7	**74.5**	**71.5**	**70.0**	**76.0**

➔ compares favorably to the state of the art on both datasets

➔ even good results with no corpus training annotations for Quaero

Auxiliary experiments

We document other findings:

- ➔ training with multilingual BERT does not improve performance vs English BERT
- ➔ English only model + machine translation < our bilingual model
- ➔ 2-step training does not degrade the performance, but reduces the training time

Quaero (2015) F1	MEDLINE	EMEA
mBERT (multilingual)	**73.7**	76.5
camemBERT (FR)	73.5	75.5
BERT (EN)	**73.7**	**76.8**

Quaero (2015) F1	MEDLINE	EMEA
Our model	**73.7**	**76.5**
with mBERT-MT	71.8	76.5
with BERT-MT	72.4	75.5

Quaero (2015) F1		MEDLINE	EMEA
~7h training	2 steps	**73.7**	**76.5**
~15h training	1 step	73.6	76.2

Auxiliary experiments

We document other findings:

- ➔ training with multilingual BERT does not improve performance vs English BERT
- ➔ English only model + machine translation < our bilingual model
- ➔ 2-step training does not degrade the performance, but reduces the training time

Quaero (2015) F1	MEDLINE	EMEA
mBERT (multilingual)	**73.7**	76.5
camemBERT (FR)	73.5	75.5
BERT (EN)	**73.7**	**76.8**

Quaero (2015) F1	MEDLINE	EMEA
Our model	**73.7**	**76.5**
with mBERT-MT	71.8	76.5
with BERT-MT	72.4	75.5

Quaero (2015) F1		MEDLINE	EMEA
~7h training	2 steps	**73.7**	**76.5**
~15h training	1 step	73.6	76.2

Auxiliary experiments

We document other findings:

- ➔ training with multilingual BERT does not improve performance vs English BERT
- ➔ English only model + machine translation < our bilingual model
- ➔ 2-step training does not degrade the performance, but reduces the training time

Quaero (2015) F1	MEDLINE	EMEA
mBERT (multilingual)	**73.7**	76.5
camemBERT (FR)	73.5	75.5
BERT (EN)	**73.7**	**76.8**

Quaero (2015) F1	MEDLINE	EMEA
Our model	**73.7**	**76.5**
with mBERT-MT	71.8	76.5
with BERT-MT	72.4	75.5

Quaero (2015) F1		MEDLINE	EMEA
~7h training	2 steps	**73.7**	**76.5**
~15h training	1 step	73.6	76.2

Experiments: monolingual / bilingual / multilingual

Mantra

Train ▼ Test ▶	ENG	SPA	FRE	GER	DUT	All
ENG	81.1	52.2	53.0	45.9	38.7	62.2
ENG+SPA	**81.9**	72.8	60.8	49.9	40.0	67.4
ENG+FRE	81.4	56.9	73.7	48.8	40.9	67.4
ENG+GER	81.8	55.5	56.6	70.9	45.2	68.3
ENG+DUT	81.4	55.7	55.1	51.7	66.1	66.6
Multilingual	81.0	**73.4**	**74.1**	**72.9**	**68.8**	**75.7**

Quaero (2015)

	MEDLINE 2015			EMEA 2015		
	Prec.	Rec.	F1	Prec.	Rec.	F1
FR synonyms only	73.8	52.8	61.5	82.4	52.8	64.4
EN synonyms only	**79.7**	45.1	57.5	**84.3**	41.0	55.1
FR + EN synonyms	78.3	**62.1**	**69.3**	82.7	**57.4**	**67.8**

Given the same set of concepts

- ➔ Bilingual > monolingual
- ➔ Multilingual > bilingual
- ➔ Similar languages have better co-training performance

Experiments: monolingual / bilingual / multilingual

Mantra

Train ▼ Test ▶	ENG	SPA	FRE	GER	DUT	All
ENG	81.1	52.2	53.0	45.9	38.7	62.2
ENG+SPA	**81.9**	72.8	60.8	49.9	40.0	67.4
ENG+FRE	81.4	56.9	73.7	48.8	40.9	67.4
ENG+GER	81.8	55.5	56.6	70.9	45.2	68.3
ENG+DUT	81.4	55.7	55.1	51.7	66.1	66.6
Multilingual	81.0	**73.4**	**74.1**	**72.9**	**68.8**	**75.7**

Quaero (2015)

	MEDLINE 2015			EMEA 2015		
	Prec.	Rec.	F1	Prec.	Rec.	F1
FR synonyms only	73.8	52.8	61.5	82.4	52.8	64.4
EN synonyms only	**79.7**	45.1	57.5	**84.3**	41.0	55.1
FR + EN synonyms	78.3	**62.1**	**69.3**	82.7	**57.4**	**67.8**

Given the same set of concepts

- ➔ Bilingual > monolingual
- ➔ Multilingual > bilingual
- ➔ Similar languages have better co-training performance

Experiments: monolingual / bilingual / multilingual

Mantra

Train ▼ Test ▶	ENG	SPA	FRE	GER	DUT	All
ENG	81.1	52.2	53.0	45.9	38.7	62.2
ENG+SPA	**81.9**	72.8	60.8	49.9	40.0	67.4
ENG+FRE	81.4	56.9	73.7	48.8	40.9	67.4
ENG+GER	81.8	55.5	56.6	70.9	45.2	68.3
ENG+DUT	81.4	55.7	55.1	51.7	66.1	66.6
Multilingual	81.0	**73.4**	**74.1**	**72.9**	**68.8**	**75.7**

Quaero (2015)

	MEDLINE 2015			EMEA 2015		
	Prec.	Rec.	F1	Prec.	Rec.	F1
FR synonyms only	73.8	52.8	61.5	82.4	52.8	64.4
EN synonyms only	**79.7**	45.1	57.5	**84.3**	41.0	55.1
FR + EN synonyms	78.3	**62.1**	**69.3**	82.7	**57.4**	**67.8**

Given the same set of concepts

- ➔ Bilingual > monolingual
- ➔ Multilingual > bilingual
- ➔ Similar languages have better co-training performance

Some examples

System	Example mention	Expected concept + synonyms	Predicted concept + synonyms
MLNorm	greffon renal	**C1261317** • [EN] transplanted kidney • [EN] kidney transplant • [EN] structure of transplanted kidney	✓
	cinquième métacarpien	**C0730166** • [EN] bone structure of fifth metacarpal • [EN] fifth metacarpal bone	✓
	vaccination par le b.c.g	**C0199804** • [FR] immunisation contre la tuberculose • [EN] bcg vaccination	✓
	in vitro	**C0681828** • [EN] in vitro study • [EN] study vitro	**C3850137** • [EN] in vitro techniques • [EN] technique in vitro • [EN] in vitro as topic
	coffea robusta	**C0678439** • [EN] coffea rubusta (food)	**C1138610** • [EN] coffea arabica
mBERT-MT	cellar (translated from the French "cave")	**C0042460** • [EN] vena cava structure • [EN] venae cavae	**C0007634** • [EN] cell • [EN] cell structure
	be careful (translated from the French "attention")	**C0004268** • [EN] attention	**C3257858** • [EN] my thinking is usually careful and purposeful

Key contributions

➔ good results even without manually annotated data using knowledge from UMLS

➔ multilingual model > bilingual model > monolingual model

➔ pre-training embeddings matter less than expected

➔ two steps training can be used to speed up training

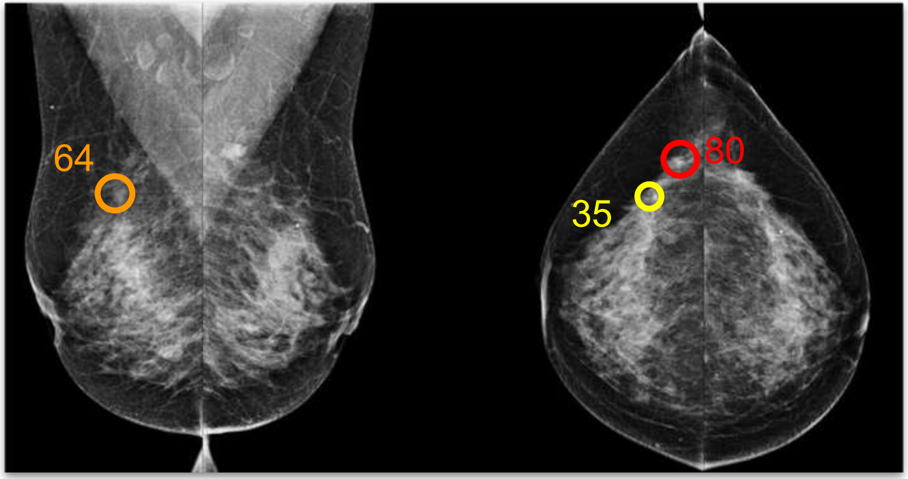
Structured entity extraction

Medical context

Prevention

➔ Breast cancer detection using machine learning models on images

➔ Validate or train using existing data: need to extract it and make it queriable



Patient follow-ups

➔ Extract patient history

➔ Detect lesion evolutions

➔ Map different reports together

Complex entities...

Frame type	Field	Field value
Cancer Risk	score trigger	
	score type	type 0 / type 1 / ... type 6
	laterality	left / right
	temporality	overlap / before doc time
Breast density	density trigger	
	density type	type 1 / type 2 / type 3 / type 4
	laterality	left / right
	temporality	overlap / before doc time
Diagnostic procedure	diag. trigger	
	diag. type	mammography / ultrasound / ...
	organ	breast / other
	laterality	left / right
	temporality	overlap / before / after doc time
Therapeutic procedure	ther. trigger	
	ther. type	surgery / other
	organ	breast / other
	laterality	left / right
	temporality	overlap / before / after doc time
Radiological lesion	lesion trigger	
	organ	breast / other
	laterality	left / right
	temporality	overlap / before doc time
	quadrant	lower inner / axillary region / ...
	size	
	distance	
	angle	

- ➔ Different kinds of structured entities
- ➔ **Multiple fields** per entity
- ➔ **Justify** each field in the text

Lesion 1	Frame 1		Frame 2	
field	**value**	**justification**	**value**	**justification**
trigger		[kystes], [nodules]		Plusieurs [kystes]
organ	breast	[mammaire]	breast	
laterality	left	[Gauche]:	left	à [gauche]
temporality	overlap		overlap	
quadrant				
size		[millimetrique]		
distance	30mm	[3 cm]		
angle	8	[8h]		

… in complex documents

INDICATION:
Previous history of breast neoplsia in the sister at 54.
Personal history: notion of surgery for breast cyst.

Results :
Breasts of symmetrical volume, density graded II.
Dystrophic calcifications scattered in the left breast.

Complementary ultrasound:
Left breast: Multiple stable homogeneous hypoechoic nodular formations are found compared to previous ultrasound, compatible with fibroadenomas located as follows:
On the 10 o'clock and 11 o'clock position at 3 cm from the nipple two nodules of 3 x 8 mm.
On the 2 o"clock pos. at 2 cm measuring 3 x 8 mm.
Apparition of a 5 x 11 mm nodule in the LI quadrant on the 8 o'clock position at 2 cm from the nipple.
Right breast: hypoechoic microformations smaller than 5mm.

CONCLUSION :
Multiple nodules of the left breast compatible with stable fibroadenomas with this day appearance of a
HospitalName
123 Main Street City Cedex
centimetric lower-inner left nodule. Further ultrasound surveillance is advised. ACR 3 for both breasts.

- Long documents
- Typos
- Ambiguous sections
- Overlapping, elliptic structures
- PDF→text artefacts
- Implicit information (e.g. time)
- …

Report annotation

→ Let’s focus on an example and extract lesion entities

Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm sur le rayon de 6h. Ces nodules sont millimétriques.
Droite:
Pas de masse suspecte.

CONCLUSION:
Plusieurs kystes à gauche.

Report annotation

➔ We fill a first frame…

Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm sur le rayon de 6h. Ces nodules sont millimétriques.
Droite:
Pas de masse suspecte.

CONCLUSION:
Plusieurs kystes à gauche.

Lesion 1	Frame 1	
field	**value**	**justification**
trigger		[kystes], [nodules]
organ	breast	[mammaire]
laterality	left	[Gauche]:
temporality	overlap	
quadrant		
size		[millimetrique]
distance		[3 cm]
angle		[8h]

Report annotation

- ➔ We fill a first frame…
- ➔ … part of an object referred in 2 places

Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm
sur le rayon de 6h. Ces nodules
sont millimétriques.
Droite:
Pas de masse suspecte.

CONCLUSION:
Plusieurs kystes à gauche.

Lesion 1	Frame 1		Frame 2	
field	**value**	**justification**	**value**	**justification**
trigger		[kystes], [nodules]		Plusieurs [kystes]
organ	breast	[mammaire]	breast	
laterality	left	[Gauche]:	left	à [gauche]
temporality	overlap		overlap	
quadrant				
size		[millimetrique]		
distance		[3 cm]		
angle		[8h]		

Report annotation

→ There is a second lesion

Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm sur le rayon de 6h. Ces nodules sont millimétriques.
Droite:
Pas de masse suspecte.

CONCLUSION:
Plusieurs kystes à gauche.

Lesion 2	Frame 3		Frame 2	
field	**value**	**justification**	**value**	**justification**
trigger		[kystes], [nodules]		Plusieurs [kystes]
organ	breast	[mammaire]	breast	
laterality	left	[Gauche]:	left	à [gauche]
temporality	overlap		overlap	
quadrant				
size		[millimetrique]		
distance		[2 cm]		
angle		[6h]		

Report annotation

- ➔ There is a second lesion
- ➔ The 2 lesions overlap in many places

Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm sur le rayon de 6h. Ces nodules sont millimétriques.
Droite:
Pas de masse suspecte.

CONCLUSION:
Plusieurs kystes à gauche.

Lesion 1	Frame 1		Frame 2	
field	**value**	**justification**	**value**	**justification**
trigger		[kystes], [nodules]		Plusieurs [kystes]
organ	breast	[mammaire]	breast	
laterality	left	[Gauche]:	left	à [gauche]
temporality	overlap		overlap	
quadrant				
size		[millimetrique]		
distance		[3 cm]		
angle		[8h]		

Lesion 2	Frame 3		Frame 2	
field	**value**	**justification**	**value**	**justification**
trigger		[kystes], [nodules]		Plusieurs [kystes]
organ	breast	[mammaire]	breast	
laterality	left	[Gauche]:	left	à [gauche]
temporality	overlap		overlap	
quadrant				
size		[millimetrique]		
distance		[2 cm]		
angle		[6h]		

Report annotation

→ We annotate other types of entities

Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm sur le rayon de 6h. Ces nodules sont millimétriques.
Droite:
Pas de masse suspecte.

CONCLUSION:
Plusieurs kystes à gauche.

Diag. proc. 1	Frame 4		Frame 5	
field	**value**	**justification**	**value**	**justification**
trigger		[Echographie]		[Echographie]
organ	breast	[mammaire]	breast	[mammaire]
laterality	left	[Gauche]:	right	[Droite]
temporality	maintenant		maintenant	
diag type	ultrasound	[Echographie]	ultrasound	[Echographie]

Lesion 1	Frame 1		Frame 2	
field	value	justification	value	justification
trigger		[kystes], [nodules]		Plusieurs [kystes]
organ	breast	[mammaire]	breast	
laterality	left	[Gauche]:	left	à [gauche]
temporality	overlap		overlap	
quadrant				
size		[millimetrique]		
distance		[3 cm]		
angle		[8h]		

Lesion 2	Frame 3		Frame 2	
field	value	justification	value	justification
trigger		[kystes], [nodules]		Plusieurs [kystes]

Annotation result

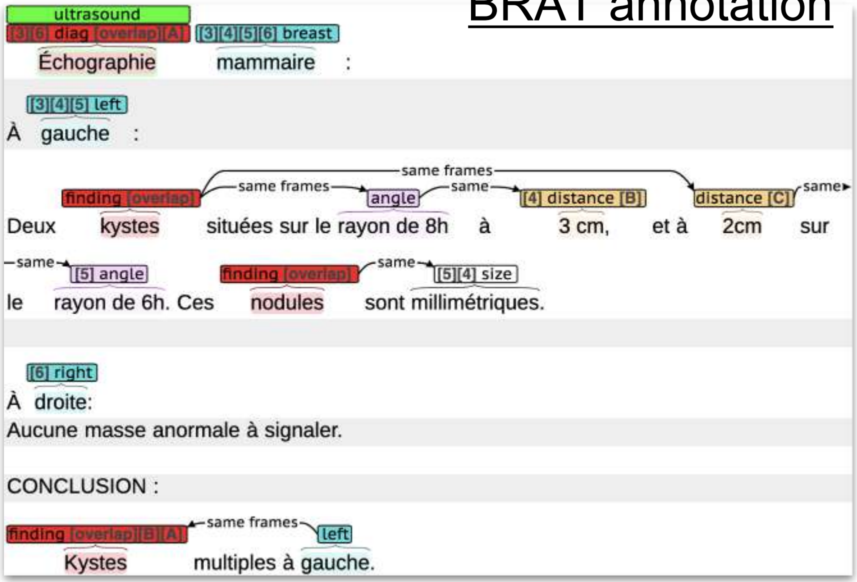
Total objects/frames:

	train		test	
	object	frame	object	frame
radiological lesion	279	449	122	210
diagnostic procedure	285	795	141	379
therapeutic procedure	51	83	22	29
BIRADS score	152	152	82	82
breast density	98	98	52	52

Per document:

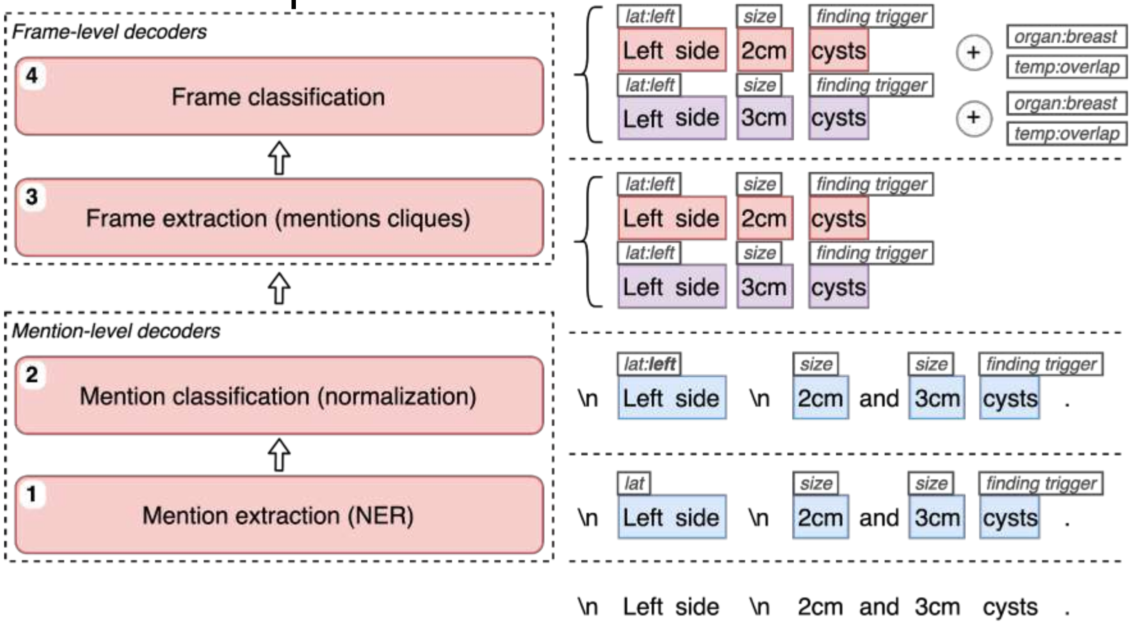
	train	test
count	80	40
average words	361.08	362.18
average lines	45.74	45.48
average frames	19.48	18.42
average objects	10.81	10.48

BRAT annotation



The method

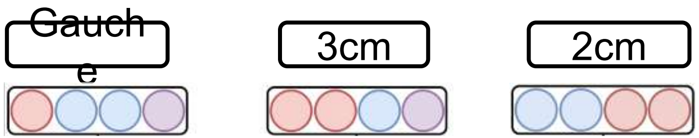
We split the problem in multiple subtasks

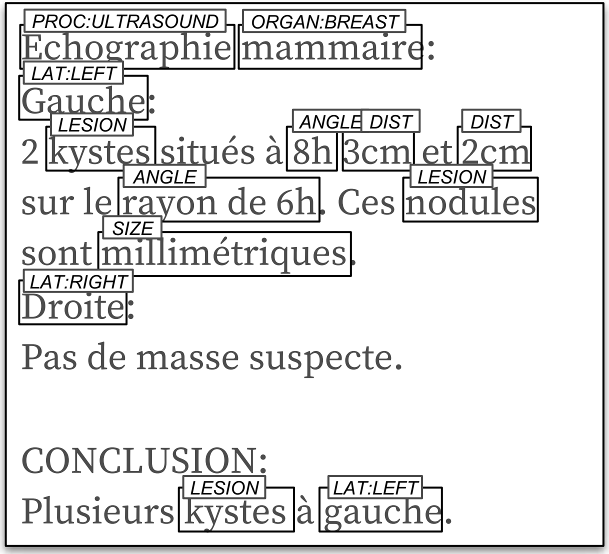


Simple bricks: normalized named entities

Build on the previously described tasks

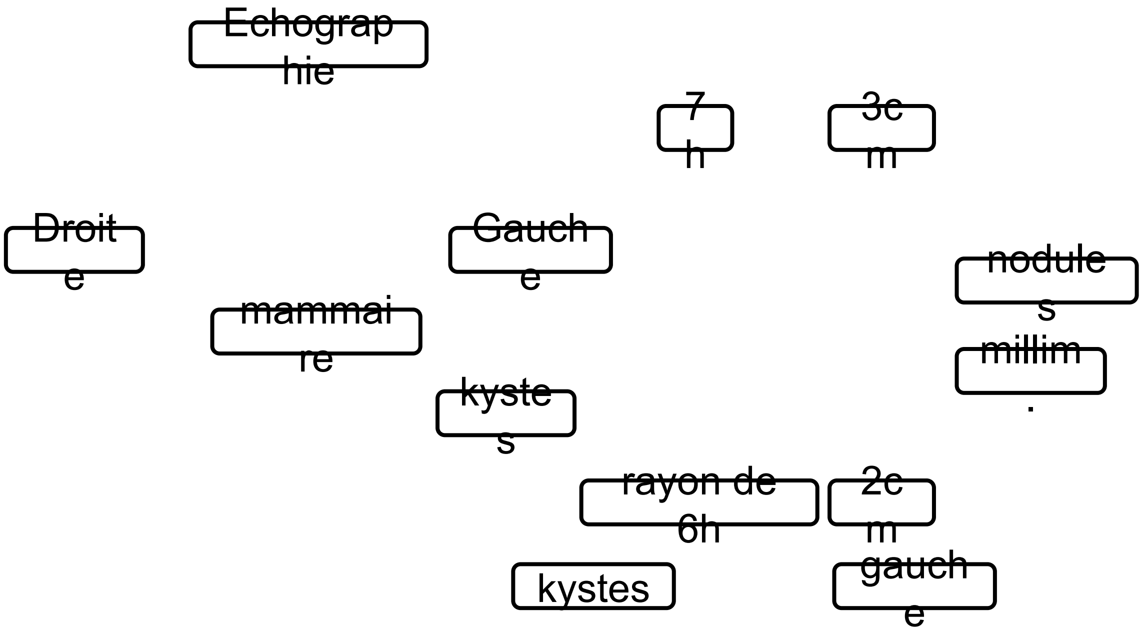
- ➔ Extract named entities using a model as described earlier
- ➔ Normalize them using a classification model (some entities may have multiple concepts)
- ➔ Compute an embedding of mentions with the average of the embeddings for each word





Frame extraction

➔ Build a graph by asking: "**Do these two entities belong to the same frame(s) ?**"

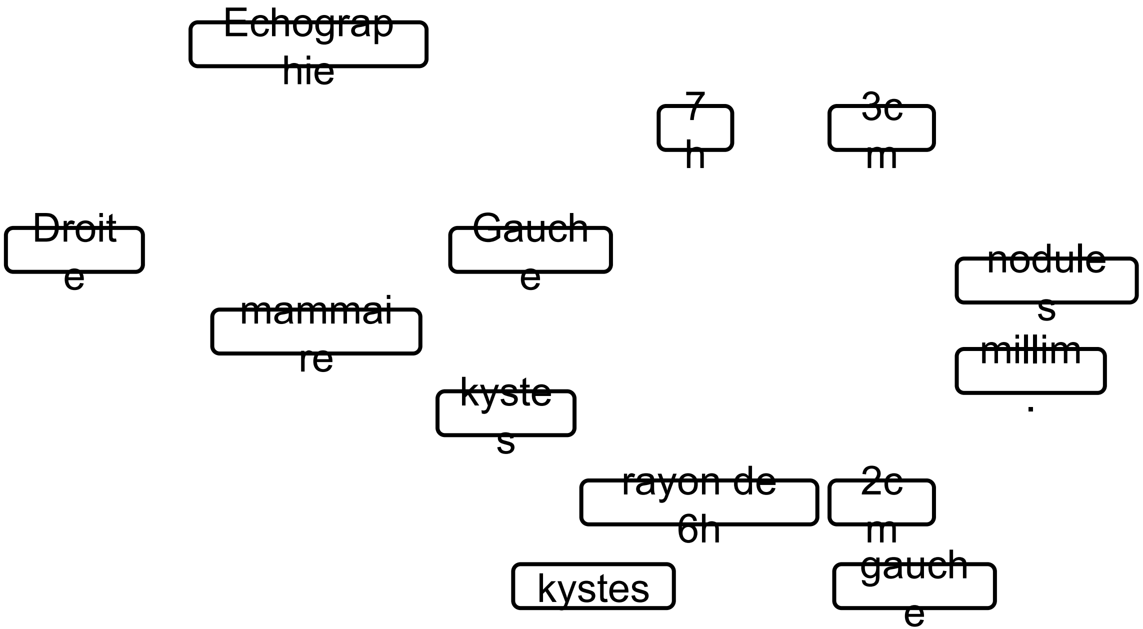


Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm sur le rayon de 6h. Ces nodules sont millimétriques.
Droite:
Pas de masse suspecte.

CONCLUSION:
Plusieurs kystes à gauche.

Frame extraction

➔ Build a graph by asking: "**Do these two entities belong to the same frame(s) ?**"

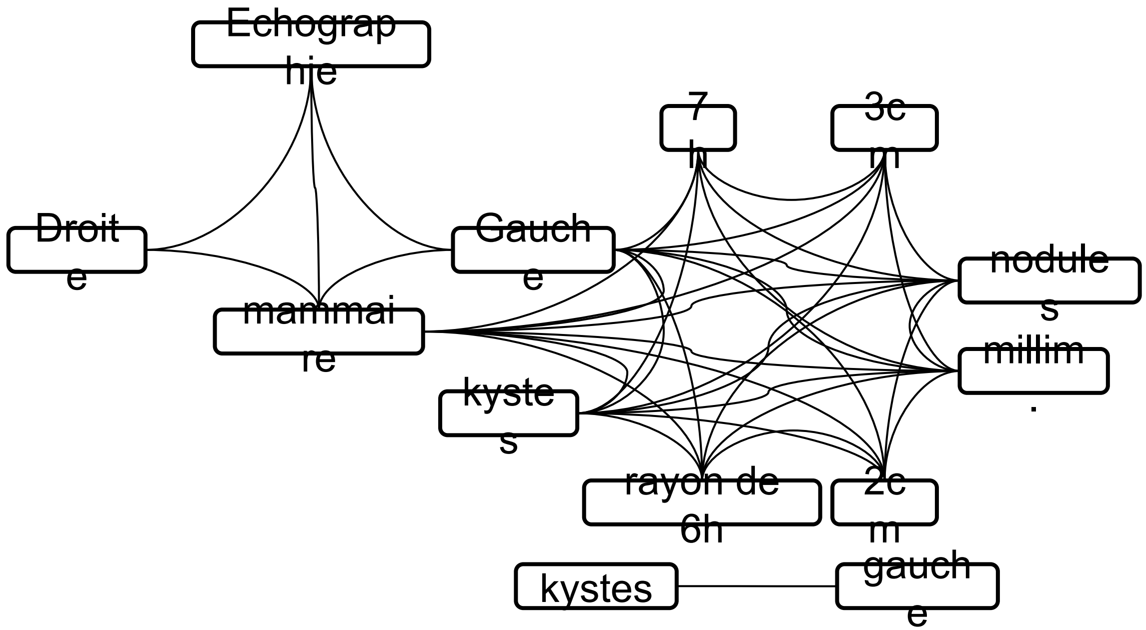


Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm sur le rayon de 6h. Ces nodules sont millimétriques.
Droite:
Pas de masse suspecte.

CONCLUSION:
Plusieurs kystes à gauche.

Frame extraction

➔ Build a graph by asking: “**Do these two entities belong to the same frame(s) ?**”



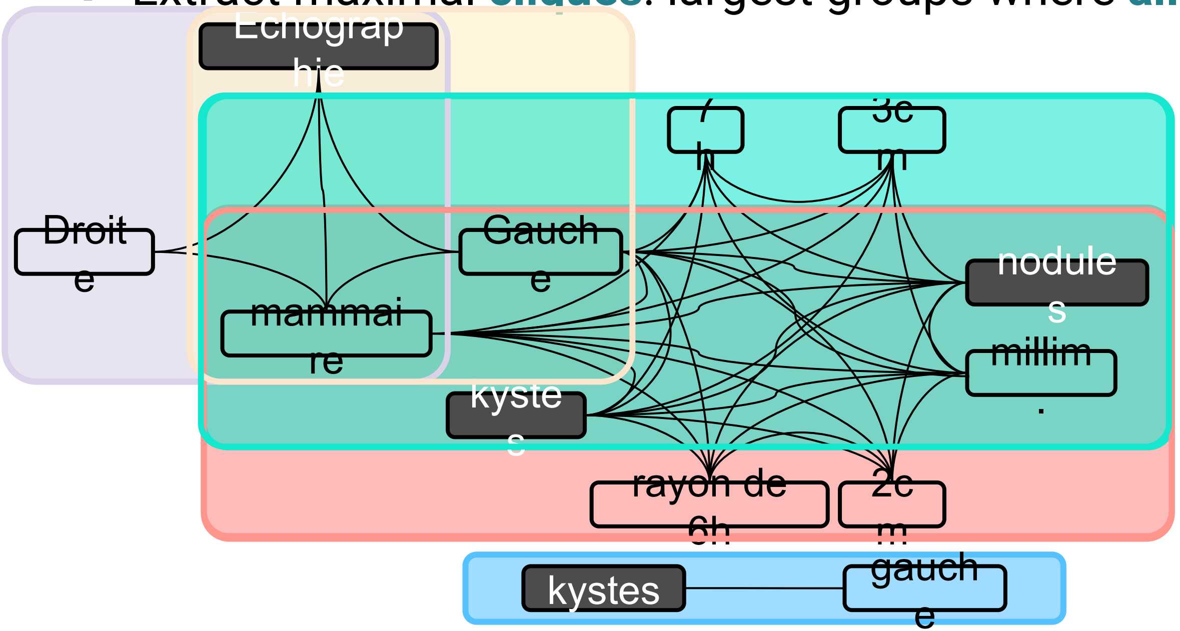
Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm sur le rayon de 6h. Ces nodules sont millimétriques.
Droite:
Pas de masse suspecte.

CONCLUSION:
Plusieurs kystes à gauche.

Frame extraction

➔ Build a graph by asking: “**Do these two entities belong to the same frame(s) ?**”

➔ Extract maximal **cliques**: largest groups where **all** entities agree with each other



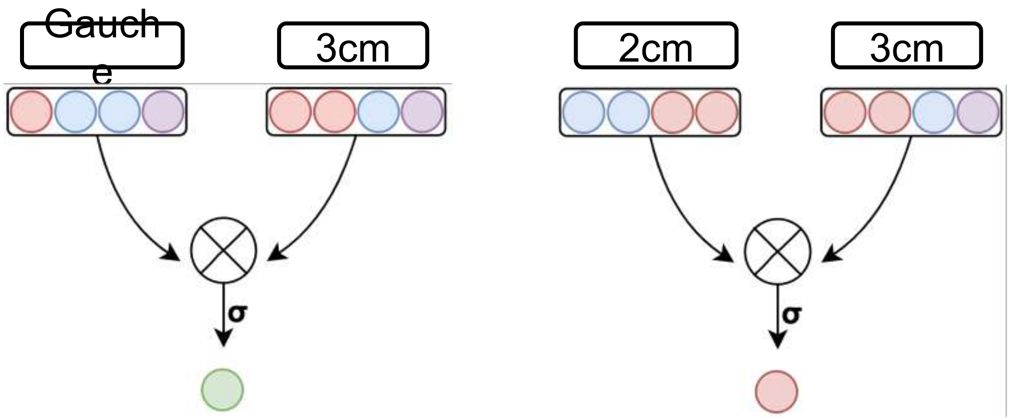
Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm sur le rayon de 6h. Ces nodules sont millimétriques.
Droite:
Pas de masse suspecte.

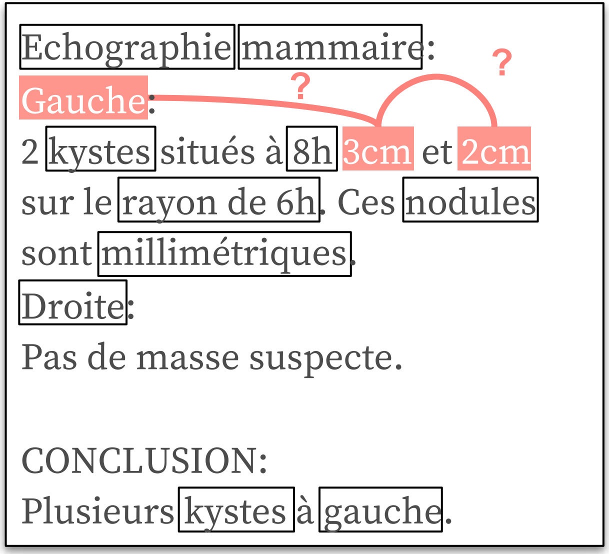
CONCLUSION:
Plusieurs kystes à gauche.

Frame extraction: simple relations

➔ How do we decide if two mentions should be linked ?

➔ Should we simply match embeddings together ?





➔ Yes, but not only

Frame extraction: scope relations

→ Concept of **scope** relations: text area does an entity convey its meaning

→ Mix “**scopes**” linking with simple “**matching**”

→ **Assymetric** relation: special training procedure

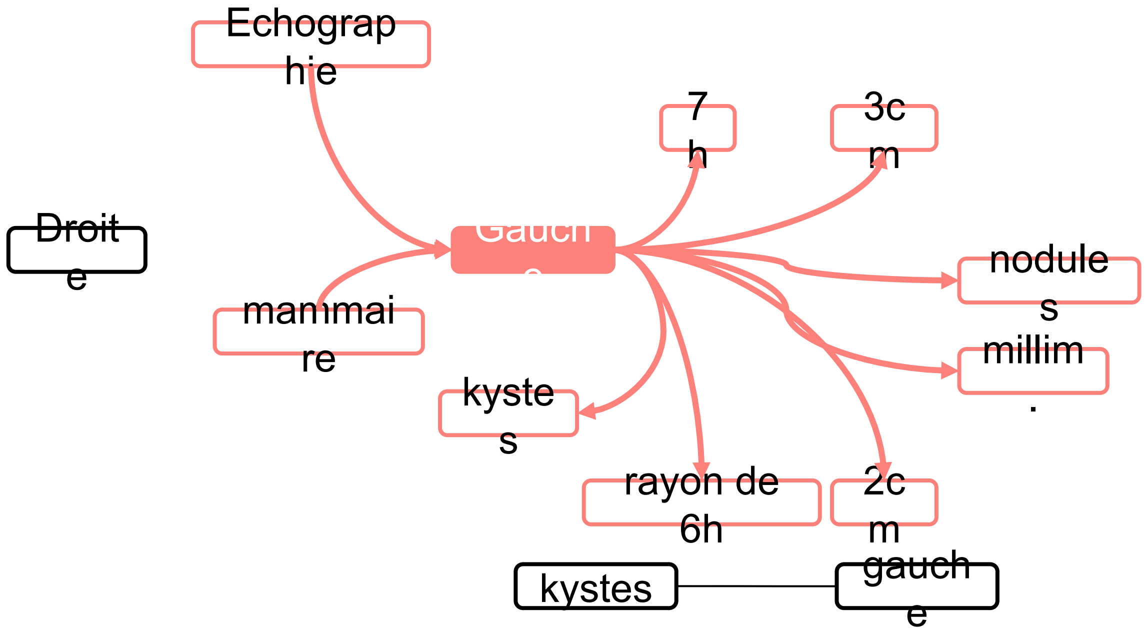
→ **Latent** scopes: the model learn the scopes on its own, since **no direct supervision** information about them: we only know which mentions should be together

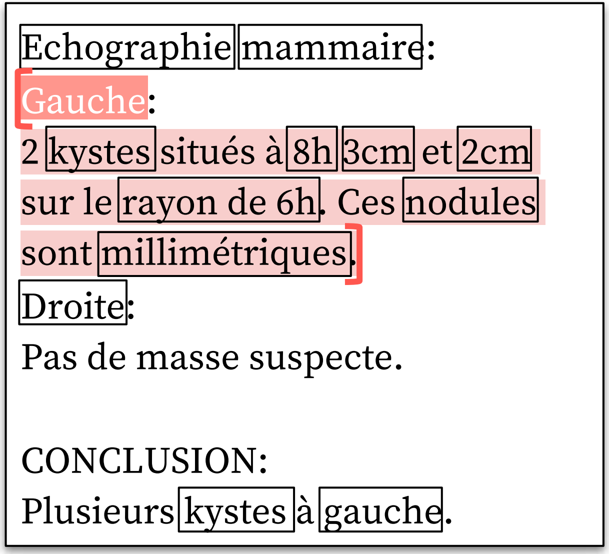
Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm sur le rayon de 6h. Ces nodules sont millimétriques.
Droite:
Pas de masse suspecte.

CONCLUSION:
Plusieurs kystes à gauche.

Frame extraction: scope relations

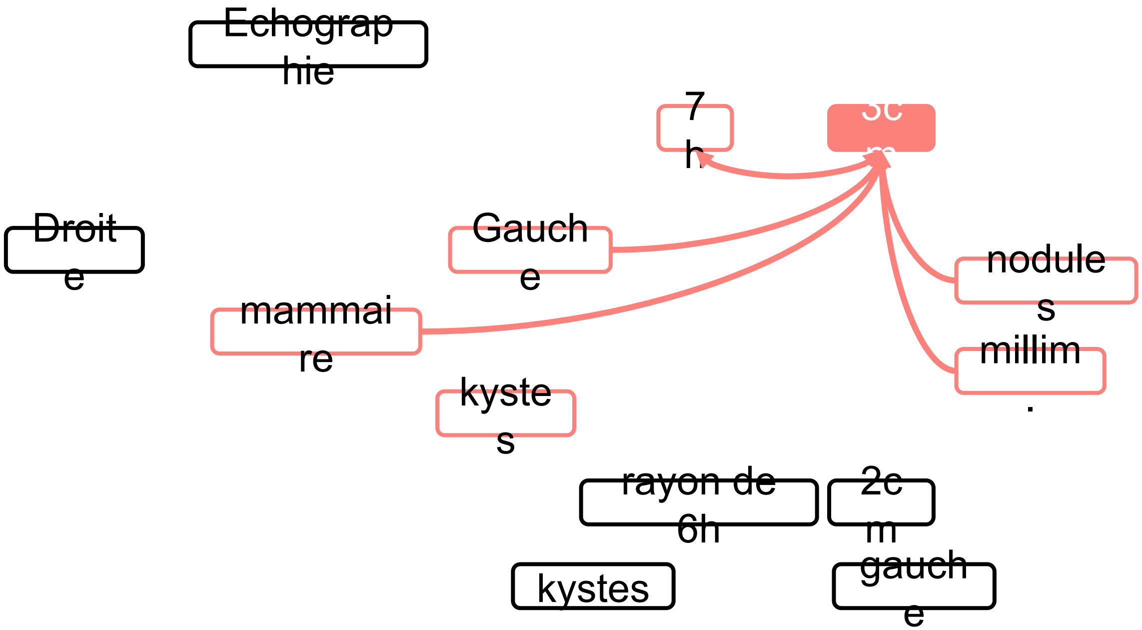
➔ Connect pairs of entities that are part of the same frame(s)





Frame extraction: scope relations

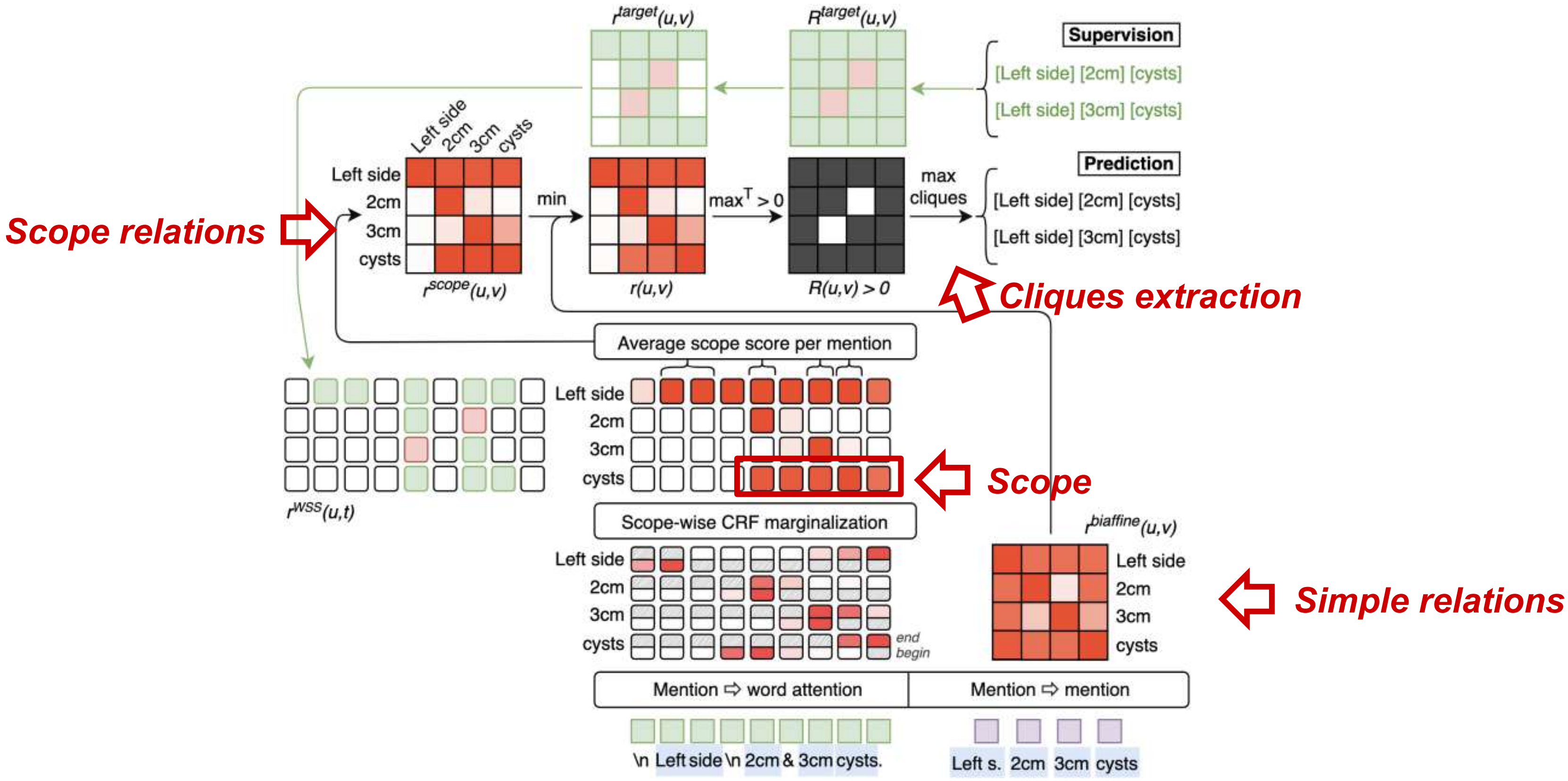
→ Connect pairs of entities that are part of the same frame(s)



Echographie mammaire:
Gauche:
2 kystes situés à 8h 3cm et 2cm
sur le rayon de 6h. Ces nodules
sont millimétriques.
Droite:
Pas de masse suspecte.

CONCLUSION:
Plusieurs kystes à gauche.

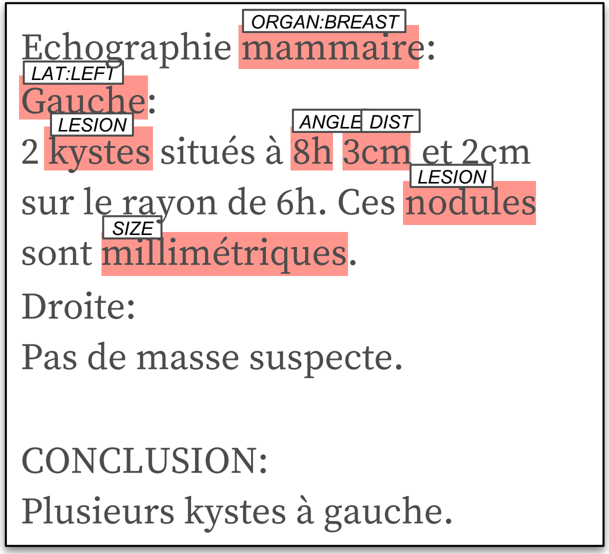
Frame extraction: the architecture



Composition: frame classification

Finally, fill in the mandatory fields that were not explicitly found in the text using a constrained classification model

Lesion 1	Frame 1	
field	**value**	**justification**
trigger		[kystes], [nodules]
organ	breast	[mammaire]
laterality	left	[Gauche]:
temporality	?	
quadrant		
size		[millimetrique]
distance		[3 cm]
angle		[8h]



Knowledge injection: synthetic sentences

➔ Automatically build synthetic dummy sentences from a small **lexicon** to **boostrap** the NER and normalization steps

➔ New sentences are mixed with the original corpus

Synonyms	*Concepts*
de [6mm]	size
en [2014]	temporality_passed
[nodule]	lesion_trigger
[mammaire]	organ_breast

...

⇨

Le concept est de **[6mm]** .

Le concept est en **[2014].**

Il y a **[nodule]**.

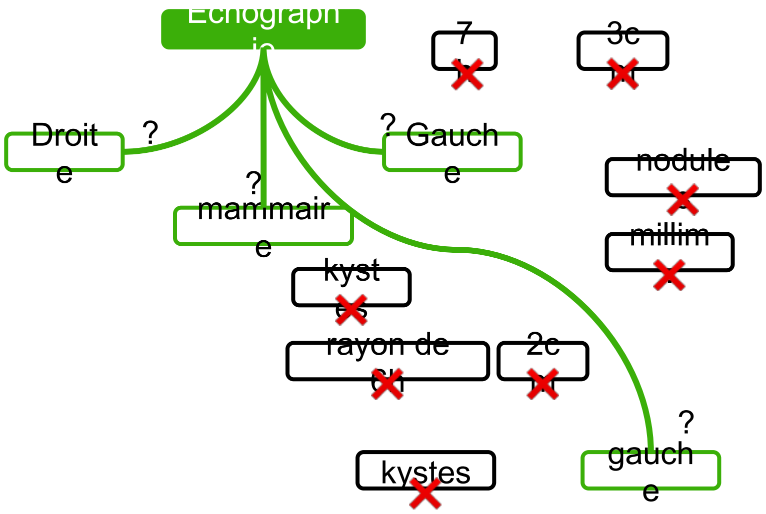
Le concept est **[mammaire]**.

...

Knowledge injection: constraints

➔ Filter **legal relations** when building the graph

➔ Filter **legal frame labels combinations** during final frame classification



Experiments: general results

- ➔ Correct performance given the small number of documents and their complexity
- ➔ Can be used to pre-annotate
- ➔ Simpler entities obtain better results

Frame type / F1	Mention Half	Frame Support	Frame Label
BIRADS score		92.5	83.3
Breast density		90.5	88.1
Diagnostic procedure		86.6	78.1
Therapeutic procedure		86.6	68.6
Lesion		78.0	62.9
Overall	96.2	85.3	72.2

Query	F1
Is mammography ?	93.9
Has passed surgery ?	73.7
Current BIRADS score	97.1
Current lateralized BIRADS score	92.0
Current breast density	92.6
Current lateralized breast density	90.5
Current lesion with quadrant	83.2
Current lesion with quadrant or radial position	77.9
Current lesion with quadrant or radial position & size	77.5

Auxiliary experiments

		Mention	Frame support	Frame label
	Base	96.2	85.3	72.2
Neural net tricks	– input-residual	95.2 (–0.9)	83.9 **(–1.4)**	69.3 **(–2.9)**
	– relative attention	95.6 (–0.5)	84.0 **(–1.3)**	70.5 **(–1.8)**
Frame extraction	– relation heuristics supervision	96.1 (–0.1)	85.4 (+0.1)	71.8 (–0.4)
	– word-level scope supervision	96.1 (–0.1)	82.1 (–3.2)	69.5 (–2.7)
	– word-level – asymmetric scope sup.	95.9 (–0.3)	74.4 (–10.9)	57.5 (–14.8)
	– scopes (only simple)	96.2 (–0.0)	80.4 (–4.9)	66.9 (–5.3)
Knowledge injection	– doc splitting (1)	96.1 (–0.0)	85.3 (+0.1)	71.5 (–0.7)
	– synthetic sentences (2)	95.4 (–0.8)	85.0 (–0.3)	70.8 (–1.5)
	– data augmentations (1+2)	95.4 (–0.8)	85.0 (–0.3)	69.9 (–2.3)
	– constraints during training	96.2 (–0.0)	84.0 (–1.3)	69.4 (–2.8)

➔ architecture matters

➔ procedure to train the model matters (especially scopes)

➔ scope relations improve the performance

➔ constraints improve the performance

➔ augmentations improve the performance

Auxiliary experiments

		Mention	Frame support	Frame label
Base		96.2	85.3	72.2
Neural net tricks	– input-residual	95.2 (–0.9)	83.9 (–1.4)	69.3 (–2.9)
	– relative attention	95.6 (–0.5)	84.0 (–1.3)	70.5 (–1.8)
Frame extraction	– relation heuristics supervision	96.1 (–0.1)	85.4 (+0.1)	71.8 (–0.4)
	– word-level scope supervision	96.1 (–0.1)	82.1 **(—3.2)**	69.5 **(—2.7)**
	– word-level – asymmetric scope sup.	95.9 (–0.3)	74.4 **(—10.9)**	57.5 **(—14.8)**
	– scopes (only simple)	96.2 (–0.0)	80.4 (–4.9)	66.9 (–5.3)
Knowledge injection	– doc splitting (1)	96.1 (–0.0)	85.3 (+0.1)	71.5 (–0.7)
	– synthetic sentences (2)	95.4 (–0.8)	85.0 (–0.3)	70.8 (–1.5)
	– data augmentations (1+2)	95.4 (–0.8)	85.0 (–0.3)	69.9 (–2.3)
	– constraints during training	96.2 (–0.0)	84.0 (–1.3)	69.4 (–2.8)

- ➔ architecture matters
- ➔ procedure to train the model matters (especially scopes)
- ➔ scope relations improve the performance
- ➔ constraints improve the performance
- ➔ augmentations improve the performance

Auxiliary experiments

		Mention	Frame support	Frame label
Base		96.2	85.3	72.2
Neural net tricks	– input-residual	95.2 (–0.9)	83.9 (–1.4)	69.3 (–2.9)
	– relative attention	95.6 (–0.5)	84.0 (–1.3)	70.5 (–1.8)
Frame extraction	– relation heuristics supervision	96.1 (–0.1)	85.4 (+0.1)	71.8 (–0.4)
	– word-level scope supervision	96.1 (–0.1)	82.1 (–3.2)	69.5 (–2.7)
	– word-level – asymmetric scope sup.	95.9 (–0.3)	74.4 (–10.9)	57.5 (–14.8)
	– scopes (only simple)	96.2 (–0.0)	80.4 **(–4.9)**	66.9 **(–5.3)**
Knowledge injection	– doc splitting (1)	96.1 (–0.0)	85.3 (+0.1)	71.5 (–0.7)
	– synthetic sentences (2)	95.4 (–0.8)	85.0 (–0.3)	70.8 (–1.5)
	– data augmentations (1+2)	95.4 (–0.8)	85.0 (–0.3)	69.9 (–2.3)
	– constraints during training	96.2 (–0.0)	84.0 (–1.3)	69.4 (–2.8)

- ➔ architecture matters
- ➔ procedure to train the model matters (especially scopes)
- ➔ scope relations improve the performance
- ➔ constraints improve the performance
- ➔ augmentations improve the performance

Auxiliary experiments

		Mention	Frame support	Frame label
Base		96.2	85.3	72.2
Neural net tricks	– input-residual	95.2 (–0.9)	83.9 (–1.4)	69.3 (–2.9)
	– relative attention	95.6 (–0.5)	84.0 (–1.3)	70.5 (–1.8)
Frame extraction	– relation heuristics supervision	96.1 (–0.1)	85.4 (+0.1)	71.8 (–0.4)
	– word-level scope supervision	96.1 (–0.1)	82.1 (–3.2)	69.5 (–2.7)
	– word-level – asymmetric scope sup.	95.9 (–0.3)	74.4 (–10.9)	57.5 (–14.8)
	– scopes (only simple)	96.2 (–0.0)	80.4 (–4.9)	66.9 (–5.3)
Knowledge injection	– doc splitting (1)	96.1 (–0.0)	85.3 (+0.1)	71.5 (–0.7)
	– synthetic sentences (2)	95.4 (–0.8)	85.0 (–0.3)	70.8 (–1.5)
	– data augmentations (1+2)	95.4 (–0.8)	85.0 (–0.3)	69.9 (–2.3)
	– constraints during training	96.2 (–0.0)	84.0 **(–1.3)**	69.4 **(–2.8)**

- ➔ architecture matters
- ➔ procedure to train the model matters (especially scopes)
- ➔ scope relations improve the performance
- ➔ constraints improve the performance
- ➔ augmentations improve the performance

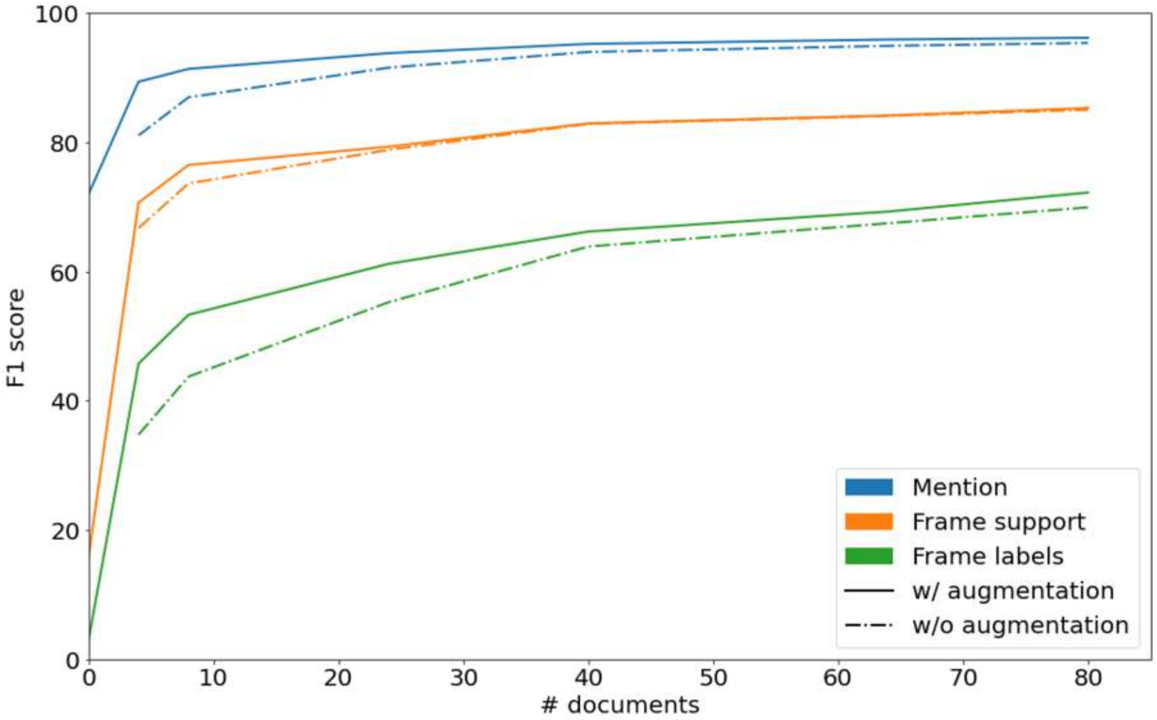
Auxiliary experiments

		Mention	Frame support	Frame label
Base		96.2	85.3	72.2
Neural net tricks	– input-residual	95.2 (–0.9)	83.9 (–1.4)	69.3 (–2.9)
	– relative attention	95.6 (–0.5)	84.0 (–1.3)	70.5 (–1.8)
Frame extraction	– relation heuristics supervision	96.1 (–0.1)	85.4 (+0.1)	71.8 (–0.4)
	– word-level scope supervision	96.1 (–0.1)	82.1 (–3.2)	69.5 (–2.7)
	– word-level – asymmetric scope sup.	95.9 (–0.3)	74.4 (–10.9)	57.5 (–14.8)
	– scopes (only simple)	96.2 (–0.0)	80.4 (–4.9)	66.9 (–5.3)
Knowledge injection	– doc splitting (1)	96.1 (–0.0)	85.3 (+0.1)	71.5 (–0.7)
	– synthetic sentences (2)	95.4 (–0.8)	85.0 (–0.3)	70.8 **(–1.5)**
	– data augmentations (1+2)	95.4 (–0.8)	85.0 (–0.3)	69.9 **(–2.3)**
	– constraints during training	96.2 (–0.0)	84.0 (–1.3)	69.4 (–2.8)

- ➔ architecture matters
- ➔ procedure to train the model matters (especially scopes)
- ➔ scope relations improve the performance
- ➔ constraints improve the performance
- ➔ augmentations improve the performance

Auxiliary experiments

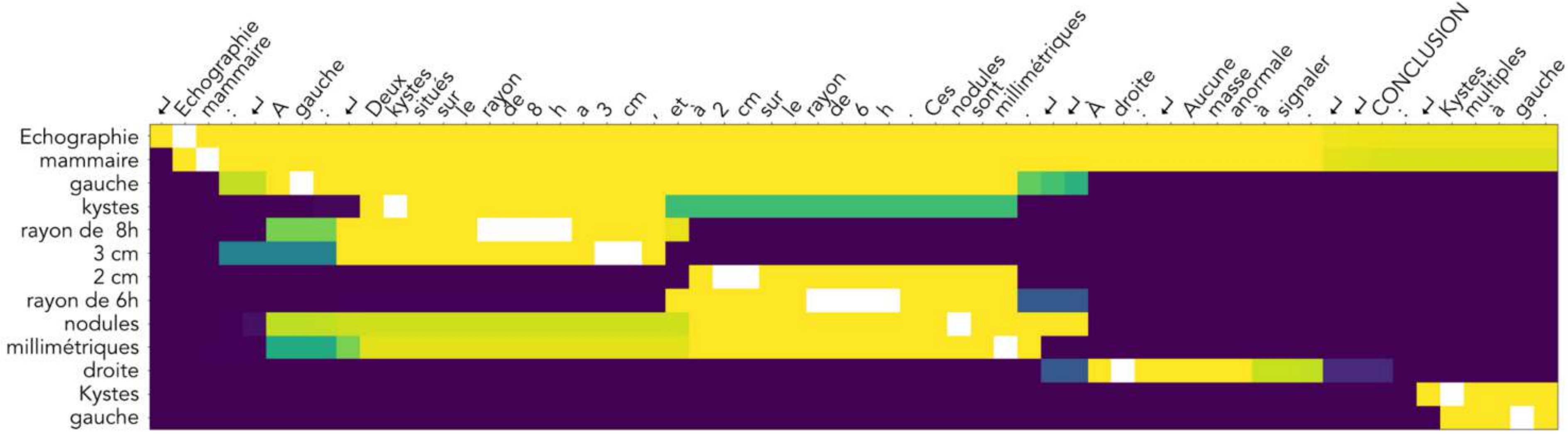
- ➔ Knowledge injection helps, especially with low number of samples
- ➔ Non zero performance with no document
- ➔ Performance improves slowly so larger number of annotations is needed



Experiments: scopes visualization

➔ Scopes capture the structure of the report

➔ Can be used to interpret results



Key findings and contributions

- ➔ Formalized a framework to structure breast imaging reports
- ➔ New dataset of breast imaging reports
- ➔ **Knowledge injection** improves the performance in a low and no data contexts
- ➔ Novel method to extract overlapping structured entities
- ➔ Concept of scope relations to improve the extraction and interpretability

Conclusion and perspectives

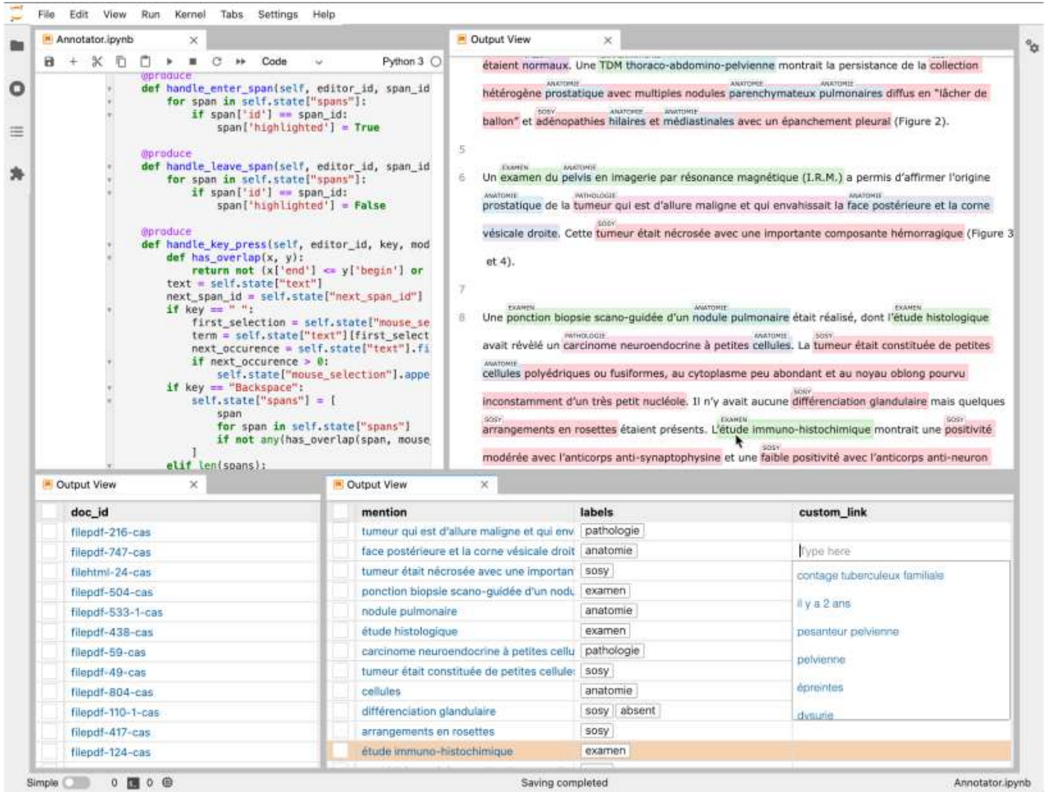
Conclusion and perspectives

- Novel methods to extract simple or structured overlapping entities in texts
- Leveraged resources in other languages and existing medical and task knowledge to bootstrap and improve our models in low data settings
- Framework and corpus to structure complex objects in radiology reports
- Proposed a novel model using scope relations and cliques to compose named entities in frames

Conclusion and perspectives

- Improve the annotation phase with custom scheme for structured data

  ⇒developed **Metanno**, a new a modulable and interactive annotation software



Conclusion and perspectives

- Connect frames together between same documents and different documents

  ⇒ same- and cross-document structured entity coreference

- Knowledge injection using constraints
  & how do extract implicit entities ? (no trigger word)

  ⇒ use first order logic frameworks and knowledge from ontologies

- Latent scopes help structuration tasks
  & large normalization training

  ⇒ pretrain embedding models built on such inductive biases and knowledge

Thank you !

Appendix

Why machine learning

Rule based models

- ➔ Need lots of handcrafting and complex rule sets
- ➔ Usually interpretable
- ➔ Need manual feature extraction, sometimes very difficult
- ➔ Do not generalize well
- ➔ Need re-engineering to improve

Machine & deep learning models

- ➔ Need lots of samples and complex architectures
- ➔ Often blackboxes
- ➔ Automatic feature extraction and capture hidden patterns
- ➔ Better generalisability (pretraining++)
- ➔ Need more corrected samples

Classic named entity recognition

- ➔ A named entity is a typed span of text
- ➔ Classically, we predict a tag for each word (e.g. using the BIOUL scheme)
- ➔ Example with two named entities in a sentence:

La patiente a un petit nodule dans le quadrant sup. ext. du sein droit .

(= The patient has a small nodule in the upper outer quadrant of the right breast.)

La	patiente	a	un	petit	nodule	dans	le	quadrant	sup.	ext.	du	sein	droit	.
⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩	⇩
∅	∅	∅	∅	∅	[]	∅	∅	[	—	—	—	—	]	∅

nodule

monocytes

Medical context