RÉSUMÉ DES EXPÉRIENCES

AGENDA

La tâche

Les données

Les expériences

Discussion

LA TÂCHE

Extraction de relations pour l'analyse des rapports de renseignement

Domaine : renseignement et défense

L'analyse de rapports est **cruciale** pour comprendre les **relations** complexes **entre** les différents **acteurs**, événements et leurs caractéristiques (entités nommées). L'extraction de relations est encore aujourd'hui un **verrou scientifique** et par conséquent nécessite un traitement manuel important. La recherche explore alors des méthodes variées afin de trouver la plus adaptée à la résolution de cette tâche.

LES DONNÉES - 1

[TRAIN] Data type : 800 entrées (id, text, entities, relations)

[TEST] Data type : 400 entrées (id, text, entities, ~~relations~~)

Exemple :

51321, "Une épidémie de dengue a fait des ravages en Equateur. Les autorités ont initié une campagne…",
[{"id": 0, "mentions": [{"value": "Equateur", "start": 45, "end": 53}], "type": "PLACE"}, etc.],
[[1, "IS_LOCATED_IN", 0], [9, "IS_IN_CONTACT_WITH", 1], etc.]

[ONTOLOGY] qui définit les classes d'entités (ACCIDENT, CIVILIAN, etc.) et les relations possibles sous format [subject_id, predicate, object_id] (ex.: [ACTOR, CREATED, ORGANISATION]

LES DONNÉES - 2

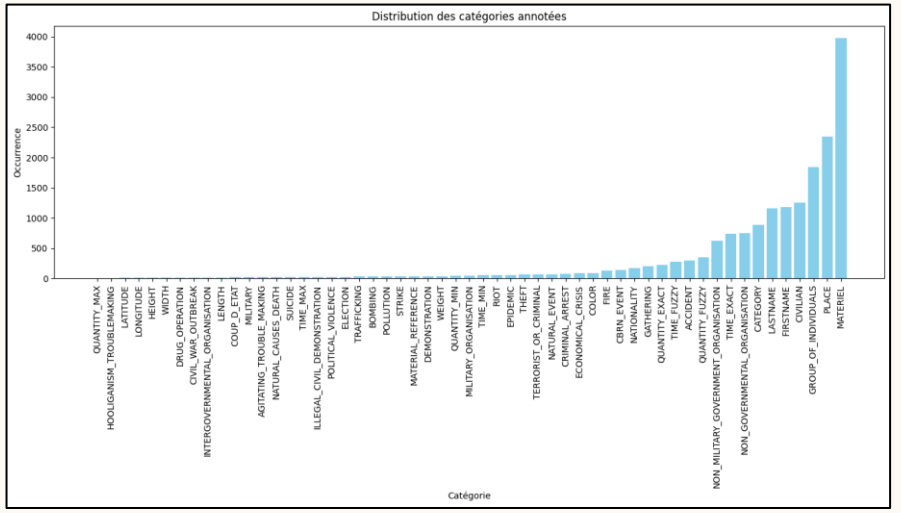
Analyse [train] :

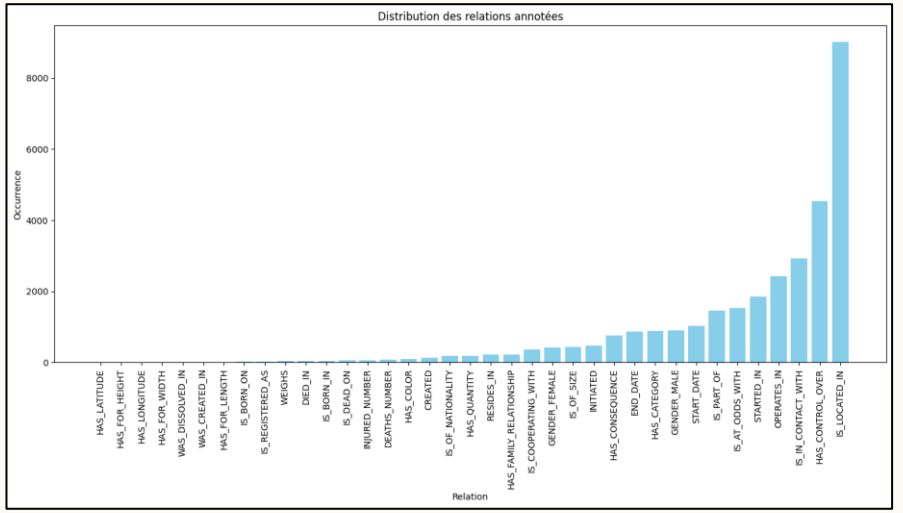
- 55 types d'entités, 17 827 occurrences
- 37 types de relations, 31 469 occurrences

Données (très) déséquilibrées, également noté par Emvista: « As shown in the previous figures **this dataset is imbalanced both with entities and relations**. Users may choose to discard low support classes."[1]

Nombre max de phrases : 15
Nombre min de phrases : 2
Nombre médian de phrases : 6.0
Nombre moyen de phrases : 6.67 (SpaCy)

Nombre moyen de tokens par texte : 244.67 (RoBERTa tokenizer)



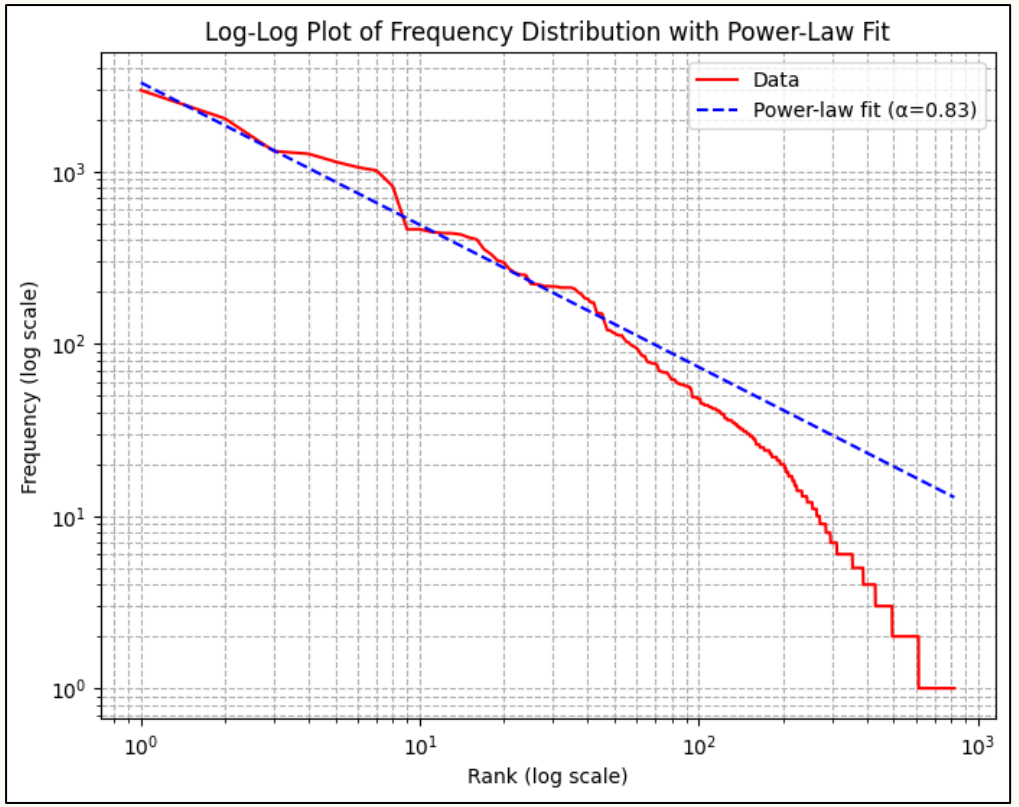


LES DONNÉES - 3

L'analyse des combinaisons [TYPE_EN_A, REL, TYPE_EN_B] nous indique une grande quantité de données très peu représentées ce qui risque de poser problème pour l'apprentissage robuste par un système.

Aucune modification pour cette expérience mais de potentiels ajustements sont possible (cf. Discussion)

Représentation naturelle de la distribution dans les rapports originaux ?



DEUX EXPÉRIENCES

Première expérience[2]

Extraire le texte autour et entre les entités afin de donner du contexte au modèle :

```
(contexte)+{marked_text[:start]}[E{entity['id']}]{value}[/E{entity['id']}]{marked_text[end:]}+(contexte)
```

```
{'text': 'arrivée sur place et les [E12]secouristes[/E12] ont déposé la [E6]victime[/E6] sur une civière. Après l’, 'label': IS_IN_CONTACT_WITH}
```

Seconde expérience

Se concentrer sur les types d’entités, sans contexte, afin d’observer si une généralisation « simple » est possible

```
{'text': 'GROUP_OF_INDIVIDUALS [SEP] PLACE', 'label’: IS_LOCATED_IN}
```

[2] inspirée de https://arxiv.org/pdf/1906.03158 et https://www.youtube.com/watch?v=luU40xu7QAg)

EXP 1 : CAMEMBERT ET N=10

Per-Class Performance Report:

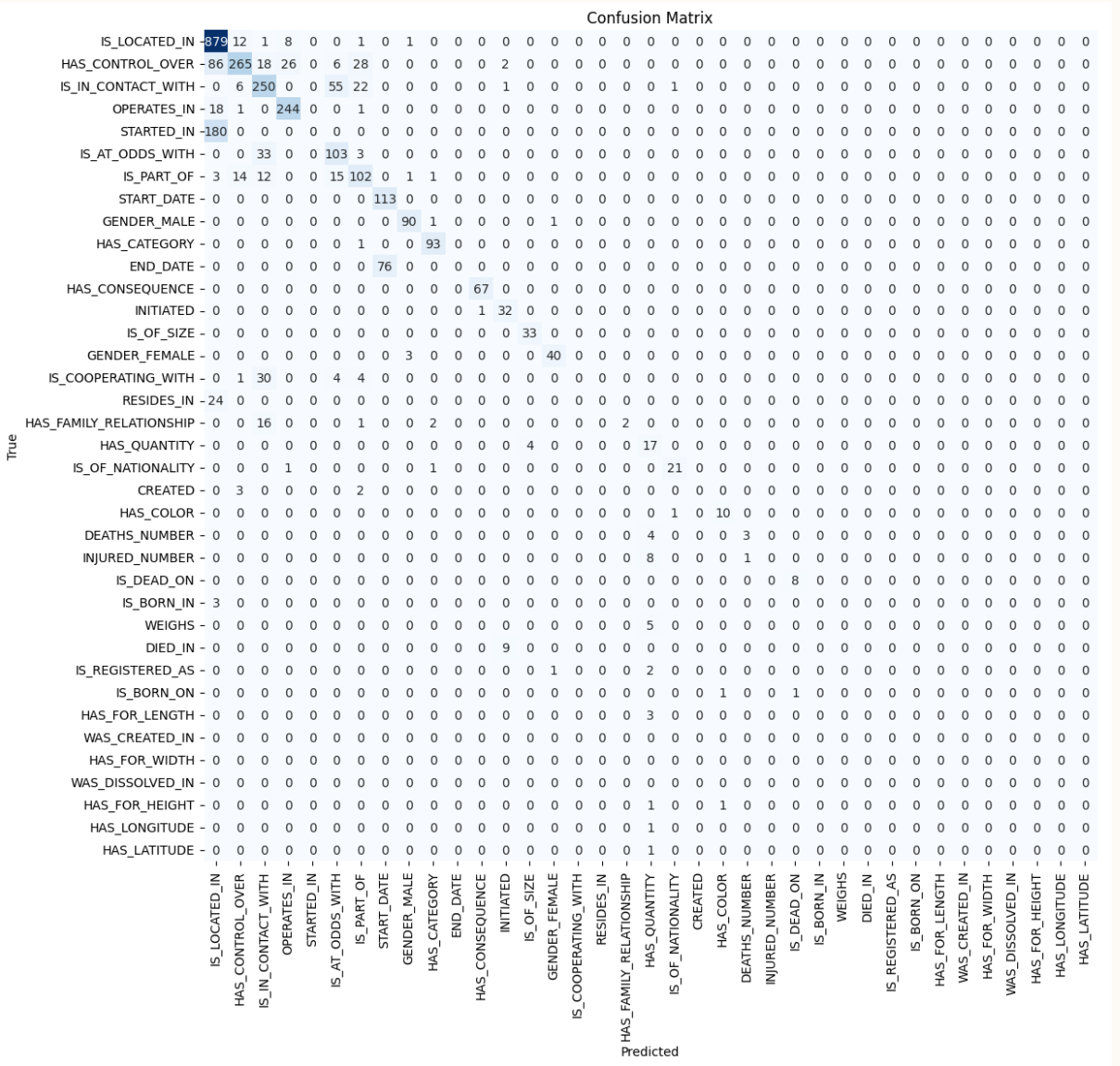
	Class	Precision	Recall	F1-Score	Support
0	IS_LOCATED_IN	0.758311	0.944072	0.841056	894
1	HAS_CONTROL_OVER	0.827893	0.670673	0.741036	416
2	IS_IN_CONTACT_WITH	0.77972	0.721683	0.749580	309
3	OPERATES_IN	0.870036	0.952569	0.909434	253
4	STARTED_IN	0.0	0.0	0.000000	186
5	IS_AT_ODDS_WITH	0.644231	0.807229	0.716578	166
6	IS_PART_OF	0.581081	0.577181	0.579125	149
7	START_DATE	1.0	1.0	1.000000	106
8	GENDER_MALE	0.990291	1.0	0.995122	102
9	HAS_CATEGORY	0.97561	1.0	0.987654	80
accuracy		-	-	0.787298	2661
macro avg		-	0.742717	0.767341	0.751958	2661
weighted avg		-	0.737305	0.787298	0.756722	2661

GENDER_MALE =
2003 à Campsas. Monsieur [E5]Guy Michel Parker[/E5][E5]Guy Michel Parker[/E5], alors ivre, a demandé à

EXP 1 : CAMEMBERT ET N=ALL

Per-Class Performance Report:

	Class	Precision	Recall	F1-Score	Support
0	IS_LOCATED_IN	0.736798	0.974501	0.839141	902
1	HAS_CONTROL_OVER	0.877483	0.614849	0.723056	431
2	IS_IN_CONTACT_WITH	0.694444	0.746269	0.719424	335
3	OPERATES_IN	0.874552	0.924242	0.898711	264
4	STARTED_IN	0.0	0.0	0.000000	180
5	IS_AT_ODDS_WITH	0.562842	0.741007	0.639752	139
6	IS_PART_OF	0.618182	0.689189	0.651757	148
7	START_DATE	0.597884	1.0	0.748344	113
8	GENDER_MALE	0.947368	0.978261	0.962567	92
9	HAS_CATEGORY	0.94898	0.989362	0.968750	94
10	END_DATE	0.0	0.0	0.000000	76
11	HAS_CONSEQUENCE	0.985294	1.0	0.992593	67
12	INITIATED	0.727273	0.969697	0.831169	33
13	IS_OF_SIZE	0.891892	1.0	0.942857	33
14	GENDER_FEMALE	0.952381	0.930233	0.941176	43
15	IS_COOPERATING_WITH	0.0	0.0	0.000000	39
16	RESIDES_IN	0.0	0.0	0.000000	24
17	HAS_FAMILY_RELATIONSHIP	1.0	0.095238	0.173913	21
18	HAS_QUANTITY	0.404762	0.809524	0.539683	21
19	IS_OF_NATIONALITY	0.913043	0.913043	0.913043	23
20	CREATED	0.0	0.0	0.000000	5
21	HAS_COLOR	0.833333	0.909091	0.869565	11
22	DEATHS_NUMBER	0.75	0.428571	0.545455	7
23	INJURED_NUMBER	0.0	0.0	0.000000	9
24	IS_DEAD_ON	0.888889	1.0	0.941176	8
25	IS_BORN_IN	0.0	0.0	0.000000	3
26	WEIGHS	0.0	0.0	0.000000	5
27	DIED_IN	0.0	0.0	0.000000	9
28	IS_REGISTERED_AS	0.0	0.0	0.000000	3
29	IS_BORN_ON	0.0	0.0	0.000000	2
30	HAS_FOR_LENGTH	0.0	0.0	0.000000	3
31	WAS_CREATED_IN	0.0	0.0	0.000000	0
32	HAS_FOR_WIDTH	0.0	0.0	0.000000	0
33	WAS_DISSOLVED_IN	0.0	0.0	0.000000	0
34	HAS_FOR_HEIGHT	0.0	0.0	0.000000	2
35	HAS_LONGITUDE	0.0	0.0	0.000000	1
36	HAS_LATITUDE	0.0	0.0	0.000000	1
accuracy		-	-	0.753734	3147
macro avg		-	0.447218	0.462149	0.436533	3147
weighted avg		-	0.683948	0.753734	0.705050	3147

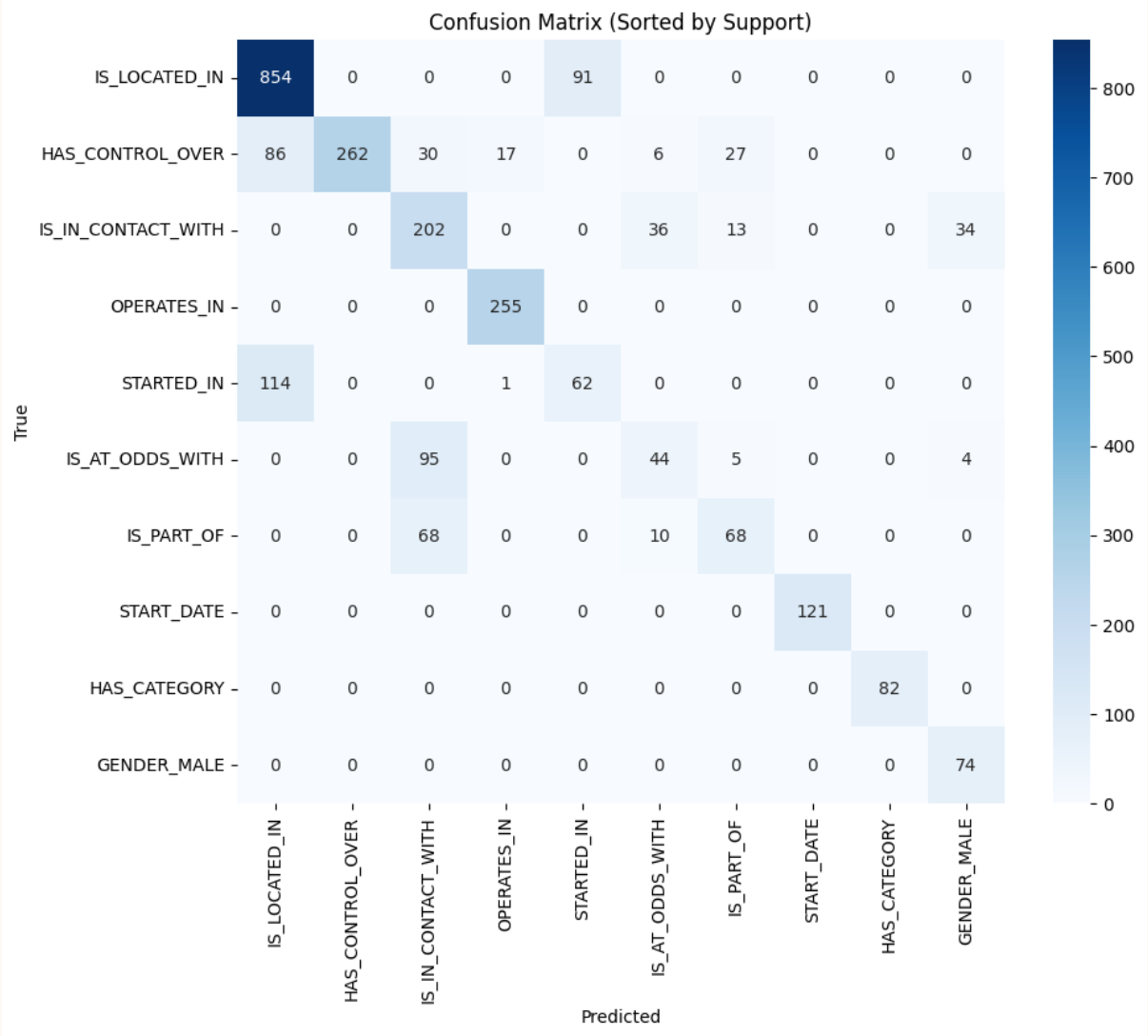


EXP 2 : DISTILBERT ET N=10

Classification Report (Sorted by Support):

	precision	recall	f1-score	support
IS_LOCATED_IN	0.81	0.90	0.85	945
HAS_CONTROL_OVER	1.00	0.61	0.76	428
IS_IN_CONTACT_WITH	0.51	0.71	0.59	285
OPERATES_IN	0.93	1.00	0.97	255
STARTED_IN	0.41	0.35	0.38	177
IS_AT_ODDS_WITH	0.46	0.30	0.36	148
IS_PART_OF	0.60	0.47	0.53	146
START_DATE	1.00	1.00	1.00	121
HAS_CATEGORY	1.00	1.00	1.00	82
GENDER_MALE	0.66	1.00	0.80	74
accuracy			0.76	2661
macro avg	0.74	0.73	0.72	2661
weighted avg	0.77	0.76	0.75	2661

{'text': 'GROUP_OF_INDIVIDUALS [SEP] PLACE’,
'label’: IS_LOCATED_IN}

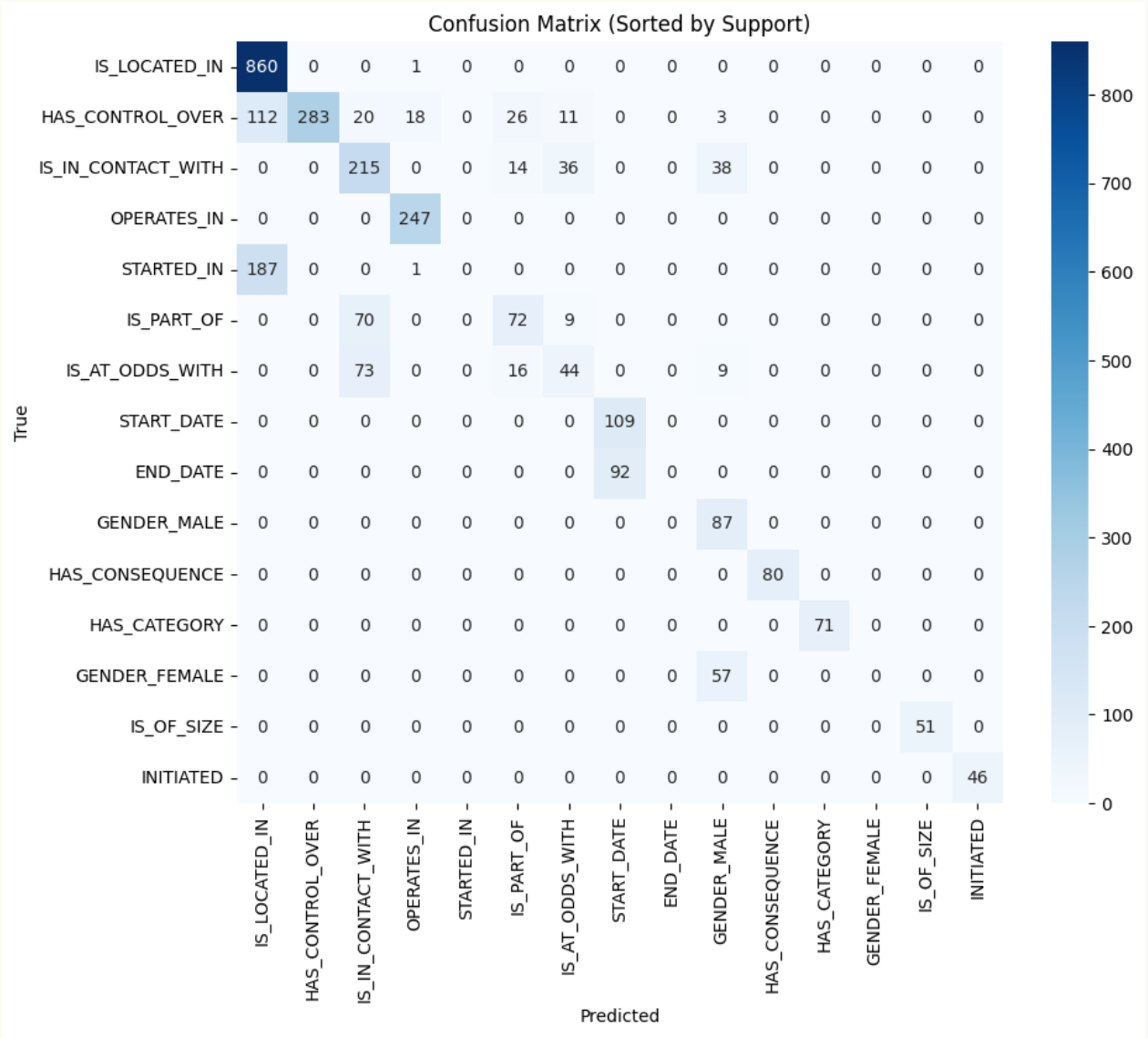


EXP 2 : DISTILBERT ET N=15

Classification Report (Sorted by Support):

	precision	recall	f1-score	support
IS_LOCATED_IN	0.74	1.00	0.85	861
HAS_CONTROL_OVER	1.00	0.60	0.75	473
IS_IN_CONTACT_WITH	0.57	0.71	0.63	303
OPERATES_IN	0.93	1.00	0.96	247
STARTED_IN	0.00	0.00	0.00	188
IS_PART_OF	0.56	0.48	0.52	151
IS_AT_ODDS_WITH	0.44	0.31	0.36	142
START_DATE	0.54	1.00	0.70	109
END_DATE	0.00	0.00	0.00	92
GENDER_MALE	0.45	1.00	0.62	87
HAS_CONSEQUENCE	1.00	1.00	1.00	80
HAS_CATEGORY	1.00	1.00	1.00	71
GENDER_FEMALE	0.00	0.00	0.00	57
IS_OF_SIZE	1.00	1.00	1.00	51
INITIATED	1.00	1.00	1.00	46
accuracy			0.73	2958
macro avg	0.62	0.67	0.63	2958
weighted avg	0.68	0.73	0.68	2958

2 epochs, 10% test



LES RÉSULTATS DES EXPÉRIENCES

- L'approche avec le contexte (expérience 1) semble la plus efficace
- Certaines classes ont des représentations trop faibles, ce qui rend leur apprentissage inefficace

Critique :

- Pas de baseline pour comparer (modèles préexistants et/ou sans fine-tuning)
- Deux modèles différents entre les deux expériences

POUR ALLER PLUS LOIN

- S'occuper des classes les plus faibles :
  - Réduire les classes populaires mais **peu de données de base** pour les classes faibles ?
  - Data augmentation en générant des phrases ?
  - Evaluation plus sévère sur les classes les plus faibles ?
  - Multiplier les epochs en risquant l'overfit ?
- Comment intégrer le système de règles de l'ontologie ? Pour quels résultats ?
- Comment prédire correctement s'il y a une relation entre deux entités ?
  (prédire pour toutes les paires semble être une mauvaise idée)
  - Commencer par associer, selon règles, les paires identifiées par ex: MATERIEL, COLOR = HAS_COLOR ?
    - Mais si plusieurs MATERIEL ou plusieurs COLOR, quelle probabilité qu'ils soient associés ?
  - Utiliser les règles syntaxiques pour la détermination de relation et la résolution de coréférence ?
    Par ex : $[\text{Marc}]_{E1a}$ est un pilote. $[\text{Il}]_{E1b}$ est membre de l'$[\text{équipe Iron}]_{E2}$.
- Comment prédire gender type (qui s'autoréférence) si le training ne fait pas de distinction de genre sur « CIVILIAN » ?
  - L'autoréférence est « résolue » en répétant l'entité dans l'expérience 1

MERCI