D'une asymétrie de corpus à une heuristique ciblée : une méthodologie LLM pour le défi TextMine - EGC 2026

Lucas Aubertin*

*Emvista
lucas.aubertin@emvista.com

Résumé. Le défi TextMine EGC 2026, proposé par la SNCF, a pour objectif la résolution du problème complexe de la désambiguïsation d'acronymes polysémiques au sein de textes réglementaires du domaine ferroviaire. Cette tâche, qui se situe à l'intersection de la reconnaissance d'entités nommées et de la liaison d'entités, revêt une importance cruciale pour l'extraction d'information en milieu industriel. Sa résolution requiert une compréhension fine du contexte afin d'associer correctement un acronyme à sa forme étendue potentiellement disponible parmi une liste de candidats fournie. Cet article présente une étude comparative des méthodologies mises en œuvre par l'auteur (équipe "Mokipo_") pour relever ce défi notamment en explorant les capacités d'un LLM sur cette tâche. Les résultats obtenus (F1-score de 88,14%, deuxième place de la compétition sur le classement privé) démontrent que, pour cette tâche de désambiguïsation spécialisée, la capacité de raisonnement contextuel des LLM, activée par des techniques de prompting avancées, surpasse de manière significative les approches de classification supervisées expérimentées.

1 Introduction

L'importance croissante du Traitement Automatique des Langues (TAL) dans des secteurs industriels spécialisés, tels que le transport ferroviaire géré par la SNCF, présente des défis singuliers. Parmi ces défis figure la gestion de la terminologie interne et plus particulièrement des acronymes. Ces derniers, bien qu'essentiels à la concision de la communication technique, présentent fréquemment un caractère polysémique ou obscur sans contexte ou connaissances extérieures pour les désambiguïser. Un même acronyme peut en effet désigner un processus, un lieu, un logiciel ou une organisation différente selon le contexte. La **résolution de cette ambiguïté** constitue un obstacle majeur pour l'indexation de documents, la construction de bases de connaissances et le partage de savoirs.

Le défi TextMine / EGC 2026 proposé par la SNCF (Lefeuvre et al., 2025) traite directement de cette problématique. La tâche consiste, pour un acronyme et un court extrait de texte l'utilisant, à identifier la (ou les) forme(s) étendue(s) correcte(s) parmi une liste d'options candidates. Le format des données (JSONL) et la métrique d'évaluation (F-mesure) définissent le cadre d'une tâche de classification binaire (mais potentiellement multi-label) pour chaque exemple proposé.

Cet article décrit la participation de l'auteur au défi TextMine 2026. Il présente l'évolution des expérimentations menées, en commençant par une approche de TF-IDF jusqu'à l'utilisation de technologies de pointes sous la forme d'un des LLM les plus performant au moment du défi (Gemini 2.5 dans sa version "Pro") (Gemini Team, 2025) en intégrant des techniques de *prompting* avancées et comparées.

La structure de cet article est la suivante : la section 2 présente un bref état de l'art sur la désambiguïsation d'acronymes. La section 3 discute des données à disposition et de leurs caractéristiques. La section 4 détaille la méthodologie et les différentes expérimentations conduites. La section 5 analyse et compare les résultats obtenus. Enfin, la section 6 discute les implications des présents travaux avant de conclure en section 7.

2 État de l'art

La désambiguïsation d'acronymes (*Acronym Disambiguation* en anglais, AD) est une tâche classique du TAL se situant à l'intersection de la reconnaissance et de la liaison d'entités (Li et al., 2018; Ciosici et al., 2019; Kugic et al., 2024). Les approches pour la résoudre ont évolué parallèlement aux paradigmes majeurs du domaine.

Historiquement, l'AD était traitée comme un problème de classification supervisée. Les méthodes fondatrices reposaient sur l'extraction de caractéristiques (par ex : vecteurs TF-IDF d'une fenêtre de contexte) et l'utilisation de classifieurs statistiques tels que les *Support Vector Machines* (SVM) (Moon et al., 2012). Si ces modèles ont établi les premières *baselines*, ils souffraient d'une incapacité à capturer la sémantique contextuelle ou la synonymie atteignant rapidement un "plafond sémantique".

L'avènement des modèles de langue pré-entraînés à base de *Transformers* tels que BERT (Devlin et al., 2019) et ses variantes françaises comme CamemBERT (Martin et al., 2020) ou plus récemment ModernCamemBERT (Antoun et al., 2025) a fait évoluer l'état de l'art. L'ajustement fin (*fine-tuning*) de ces modèles sur des données de domaine a démontré une capacité supérieure à lever les limites du "plafond sémantique". La tâche est alors typiquement formulée comme une classification binaire de paires : le modèle prend en entrée la concaténation de l'acronyme et son contexte puis une forme étendue candidate et prédit si l'association est correcte (Pan et al., 2021). Cependant, cette approche reste dépendante de la disponibilité de larges corpus annotés, souvent difficilement accessibles en milieu industriel spécialisé.

Plus récemment, l'émergence des grands modèles de langue (LLM) a introduit le paradigme de l'apprentissage en contexte (*In-Context Learning*, ICL) permettant de réaliser des tâches complexes guidées par des instructions (*prompt*) avec un nombre limité d'exemples, voire aucun (*few-shot* ou *zero-shot*), sans nécessiter de mise à jour des poids du modèle (Brown et al., 2020). Cette approche s'est récemment révélée très efficace pour la tâche de désambiguïsation d'acronymes, en particulier dans des domaines de niche comme le secteur biomédical. Des études récentes ont en effet démontré que des LLM (GPT-4 par ex.) peuvent atteindre une très haute précision en utilisant de simples instructions pour identifier la forme étendue cor-

recte d'un acronyme dans un contexte donné (Kugic et al., 2024; Vogel, 2025).

Pour augmenter la fiabilité des LLM sur des problèmes nécessitant un raisonnement plus complexe, la technique de la Chaîne de Pensée (*Chain-of-Thought*, CoT) a été développée (Wei et al., 2023). Elle améliore la performance en forçant le modèle à expliciter les étapes logiques intermédiaires qui le mènent à sa réponse finale. Cette méthode a été évaluée pour la désambiguïsation d'entités nommées homonymes dans des graphes de connaissance académique, une tâche similaire à la nôtre (Liu et Fang, 2024). Leurs travaux ont montré que l'utilisation de la CoT améliore les capacités de raisonnement et la performance du LLM par rapport aux approches ICL standards (*zero-shot* et *few-shot*).

Pour les tâches nécessitant des connaissances factuelles spécifiques, absentes ou obsolètes chez le LLM, les architectures de Génération Augmentée par Récupération (*Retrieval-Augmented Generation*, RAG) ont été proposées (Ding et al., 2024). Celles-ci ancrent la génération du modèle sur des informations extraites d'une base de connaissances externe (Lewis et al., 2020; Siriwardhana et al., 2023). Toutefois, des recherches récentes suggèrent que les architectures RAG génériques, optimisées pour des documents non structurés, sont peu adaptées à la récupération d'informations depuis des bases de connaissances structurées, comme des bases terminologiques. Pour résoudre ce problème, des approches spécialisées telles que la "*Terminology Augmented Generation*" (TAG) ont été proposées (Lackner et al., 2025). L'efficacité de cette approche repose sur l'apport au LLM de ressources lexicales structurées (par exemple les définitions d'un dictionnaire) pour guider son choix parmi les différents sens possibles d'un terme ambigu. De plus, des travaux sur la désambiguïsation en contexte RAG ont démontré la pertinence d'utiliser des ressources lexicales structurées pour fournir au LLM les différentes définitions possibles d'un terme ambigu afin de guider son choix (David et al., 2024).

Notre travail s'inscrit dans cette progression. Nous comparons les approches supervisées classiques (TF-IDF, *fine-tuning* de CamemBERT) à des stratégies basées sur les LLM (ICL + CoT) puis nous explorons une méthode hybride injectant des connaissances statiques issues d'un lexique métier, évaluant ainsi la pertinence de chaque paradigme pour le défi proposé.

3 Analyse des données à disposition

Les organisateurs ont fourni un jeu d'entraînement et un jeu de test au format JSONL en français. Une analyse statistique approfondie de ces deux jeux a été conduite pour identifier les défis de la tâche et guider notre stratégie de modélisation.

3.1 Distribution et complexité des corpus

Les jeux d'entraînement et de test présentent des tailles comparables (492 et 519 exemples respectivement) mais des distributions de données fondamentalement différentes.

— **Taille et options :** Le jeu d'entraînement contient 492 exemples portant sur 77 acronymes uniques avec une moyenne de 4,42 options par question (Max : 13, Min : 2).

Le jeu de test contient 519 exemples mais couvre une distribution plus large de 138 acronymes uniques avec une moyenne de 4,30 options (Max : 15, Min : 2).

— **Déséquilibre des acronymes :** Les deux ensembles sont fortement déséquilibrés mais pas de la même manière.
  — Le **jeu d'entraînement** est dominé par les acronymes « EF » (Entreprise Ferroviaire, 42 occurrences), « AGC » (Autorail Grande Capacité, 37) et « CLE » (Consigne Locale d'Exploitation, 33).
  — Le **jeu de test**, en revanche, est dominé par un ensemble différent d'acronymes : « RFN » (Réseau Ferré National, 68 occurrences), « TGV » (Train à Grande Vitesse, 49), « UM » (Unité Multiple, 34) et « Z2N » (Rame à deux niveaux, 31).

Cette asymétrie de distribution entre l'entraînement et le test est une observation qui suggère qu'un modèle se basant uniquement sur les statistiques apprises du jeu d'entraînement serait intrinsèquement désavantagé et indique la potentielle nécessité d'injecter des connaissances externes (via un modèle pré-entraîné sur le domaine et/ou un lexique).

3.2 Analyse de la structure des labels (jeu d'entraînement)

L'analyse des labels dans le jeu d'entraînement a révélé deux défis majeurs qui façonnent la nature de la tâche.

— **La tâche de rejet :** Une part importante des exemples d'entraînement (**13.82%** soit 68 cas sur 492) ne possède **aucune** réponse correcte. La tâche n'est donc pas seulement une classification à choix multiple, elle exige également une capacité de rejet (c'est-à-dire prédire l'absence de bonne réponse dans la liste proposée).

— **L'ambiguïté multi-label :** Une petite partie des exemples (**1.83%** soit 9 cas) possède plusieurs réponses correctes.

3.3 Identification d'un biais de construction systématique

L'analyse de ces 9 cas multi-label a révélé un indice faible mais statistiquement significatif. Sur ces 9 cas, **2 exemples** présentaient un "doublon sémantique" : une option "courte" ("Block Automatique Lumineux") et une option "longue" ("Block Automatique Lumineux : signalisation..."). Ces deux options étant toutes deux marquées comme correctes et sont identiques pour les deux exemples.

Bien que numériquement faible, ce schéma nous a incités à investiguer sa prévalence dans le jeu de test. Ne pouvant analyser les labels, nous avons recherché le *pattern* d'options lui-même (une option étant le début **exact** d'une autre avec un ratio de longueur supérieur au double). L'analyse a révélé que ce pattern n'était pas un cas isolé mais un **biais de construction systématique** du jeu de test : **109 exemples** (soit **21,00%** du jeu de test) contiennent ce pattern "court/long". Par ailleurs, ce biais est concentré sur seulement **4 acronymes** (RFN, BAL, PC, UM), l'acronyme RFN (le plus fréquent du test) étant le principal contributeur. La figure 1 pour l'acronyme "RFN" (ID : 7, jeu de test) illustre ce cas.

3.4 Présence de balises HTML

L'analyse des données révèle également des entrées pour lesquelles certaines des options contiennent des balises HTML (cf. Figure 2). Pour le jeu d'entrainement, il s'agit de 37 entrées

```
"id": 7,
"text": "d'une ou plusieurs sections de ligne du RFN. Code ligne Identifiant d'une section de ligne en combinaison avec les
valeurs kilométriques des extrémités considérées avec une précision métrique. Engin moteur terme général désignant tout
véhicule ayant la propriété de se déplacer par ses propres moyens : machine, automoteur, draisine, engin spécial motorisé non",
"acronym": "RFN",
"options": [
    "Régulation Ferroviaire Numérique",
    "Réseau Ferré National",
    "Réseau Ferré National.Ensemble des lignes ferroviaires dont la propriété et la gestion ont été confiées à SNCF Réseau par
    la loi et dont la consistance et les caractéristiques principales sont précisées par voie réglementaire.",
    "Référentiel Fonctionnel National"
]
```


FIG. 1 – *Exemple de doublon sémantique (forme courte et longue) pour "RFN".*

(soit 7,52%) concernées avec au moins une proposition contaminée. Pour celui de test, il s'agit de 18 entrées (soit 3,74%).


```
"text": "désignées dans les RT dont les rampes caractéristiques sont supérieures à 35 mm/m. Historique X76500 compatibilité :
OUI/restriction : les AGC sont ",
"acronym": "AGC",
"options": {
    "<span style='font-size:11.0pt'><span>Automoteur de Grande Capacité, construit par Bombardier. Famille d'automoteurs
    articulés de la SNCF dédiée au trafic TER et apte à 160 km/h. Cette famille est composée de 4 séries : - XGC : X 76500,
    version thermique (diesel) - ZGC : Z 27500, version électrique bi-tension (1,5 kV continu, 25 kV alternatif) - BGC : B
    81500, version bi-mode (thermique et électrique 1,5 kV continu) - Bibi ou BBGC : B 82500, version bi-mode (thermique et
    électrique) et bi-tension (1,5 kV continu, 25 kV alternatif).</span></span>": false,
    "Accord européen sur les Grandes lignes de Chemin de fer du 31 mai 1985": false,
    "Autorail Grande Capacité  ": true,
    "ARGENTON SUR CREUSE": false
}
```


FIG. 2 – *Exemple de contamination HTML pour l'acronyme "AGC" dans le jeu d'entrainement.*

3.5 Mauvaise forme pour un acronyme ?

Dans cet exemple du jeu d'apprentissage (Figure 3), il faut attribuer la bonne option à l'acronyme "RC". La seule indiquée en *True* est "Référencement Réseau" dont l'acronyme devrait être "RR" et non "RC". Sans expertise métier, il nous est difficile d'affirmer si c'est une erreur ou une particularité pour cette forme.


```
"text": "attelages (85 tonnes), avec attelages renforcés (135 tonnes), pour les trains entiers,…]. 1.3. Résumé
des modifications Indication de la mention : « La composition des catégories prédéterminées doit respecter les
masses remorquées fixées par la recommandation RC A-B 7a n°1 ». ", "acronym": "RC", "options": {"Réservoir de
Commande": false, "Renfort Clientèle": false, "Résistance de continuité": false, "Référencement Réseau": true,
"ROCHEFORT": false, "Rédacteur Consigne": false, "Raccordement Circulaire": false, "Responsabilité civile":
false, "Redevance de circulation": false, "Remise en Conformité": false, "Retour chantier ": false, "Repos
compensateur": false, "Roue codeuse": false}
```


FIG. 3 – *Exemple d'une potentielle erreur pour l'acronyme "RC" dans le jeu d'entrainement.*

4 Méthodologie et expérimentations

Cette section décrit la progression des expérimentations menées depuis les modèles initiaux basés sur le *fine-tuning* jusqu'à l'approche finale fondée sur un LLM.

4.1 Expérimentation de départ (*baseline*)

Une première expérimentation a consisté à établir un score de référence (*baseline*) en utilisant une méthode éprouvée en TAL.

Expérience 1 : *Baseline* (TF-IDF + Régression logistique). Cette approche constitue notre *baseline* pour cette étude. Techniquement, le modèle vectorise une concaténation de l'acronyme, du contexte et de l'option candidate en caractéristiques TF-IDF puis applique une régression logistique avec pondération des classes pour la classification binaire. Le F1-score obtenu est de **48,69%**.

4.2 Approche par modèles supervisés et *fine-tuning*

Afin de dépasser la performance de la *baseline*, nous avons exploré une série d'expérimentations focalisées sur l'ajustement fin (*fine-tuning*) de modèles de langue français.

4.2.1 Expérience 2 : *Fine-tuning* de CamemBERT-base

Pour dépasser la *baseline* statistique, nous avons employé CamemBERT-base (Martin et al., 2020), un modèle *Transformer* de référence pour le français en reformulant le problème en classification binaire afin d'évaluer la pertinence de chaque paire (contexte, option). Le problème est formulé par l'entrée `[CLS] acronyme+contexte [SEP] option [SEP]` et la prédiction finale a été obtenue en moyennant les probabilités issues d'une validation croisée stratifiée (5-fold) avant d'appliquer un seuil de décision (0.5). F1-score maximal obtenu : **49,33%**.

4.2.2 Expériences 3 et 4 : *Fine-tuning* optimisé de ModernCamemBERT (Gold + Silver Data)

Face au score moyen obtenu par l'expérience 2 et l'analyse quantitative des données en section 3, nous avons émis l'hypothèse qu'augmenter le volume de données d'entraînement était nécessaire. Nous avons généré un large corpus de "silver data". Cette génération a consisté à auto-annoter l'intégralité des exemples du jeu de test en utilisant l'API Gemini 2.5 Pro, qui, guidé par un *prompt few-shot* optimisé, devait déterminer la validité (*true/false*) de chaque option et la retourner au format JSON.

- **Expérience 3 :** Modèle amélioré en deux phases : un premier ajustement fin sur les données "gold" (jeu d'entrainement) suivi d'une seconde phase (avec *learning rate* abaissé) sur les données "silver". (F1-score : 78,26%)
- **Expérience 4 :** Modèle amélioré en deux phases : un premier ajustement fin sur les données "gold" (jeu d'entrainement, suivi d'une seconde phase sur un mélange de don-

nées "gold" et "silver" avec une moyenne des probabilités sur *k-fold* (k=5). (F1-score : 79,21%)

Des expérimentations complémentaires ont été menées avec un *ensembling k-fold* (k=10), en moyennant les probabilités de sortie des différents folds pour stabiliser la prédiction, en excluant le pli le plus faible et optimisant automatiquement le seuil sur le jeu de développement gold ou en utilisant un seuil par défaut de 0.5 sans réussir à surpasser le score de l'expérience 4. Cette série d'expérimentations par **modèles supervisés et *fine-tuning*** a considérablement amélioré le score qui semblait toutefois atteindre un plafond avoisinant 80% de F1-score. Nous notons cependant que la qualité de la génération des données "silver" fut un facteur déterminant pour les scores obtenus pour cette approche.

4.3 Approche par LLM

Afin de dépasser le score obtenu par la précédente approche (4.2) et en se référant à notre analyse des données (section 3) qui a révélé une asymétrie de distribution fondamentale entre les corpus d'entraînement et de test, nous avons décidé d'utiliser un type de modèle possédant une connaissance générale, externe à notre jeu de données. Dans ce but, nous avons exploré l'utilisation du LLM Gemini 2.5 Pro via son API (*batch*). Ce choix a été motivé par ses performances reconnues sur des tâches de raisonnement complexes (Gemini Team, 2025). Une température d'inférence de 0 a été employée afin d'assurer une sortie la plus déterministe possible.

4.3.1 Expérience 5 : CoT + sélection manuelle d'exemples

Pour cette expérience, le *prompt* contenait une sélection manuelle d'exemples (*few-shot*) conçue pour refléter diverses situations : 1) Une option vraie avec un contexte clair ; 2) Une option vraie nécessitant une connaissance du domaine ; 3) Une option vraie avec contexte immédiat ; 4) Aucune option vraie. Pour chaque exemple, nous avons ajouté une section indiquant le raisonnement explicite pour parvenir à la bonne réponse (ou son absence) dans le but de guider la Chaîne de Pensée (CoT) (un exemple est disponible en Annexe A.1). F1-score maximum obtenu : **83,76%**.

4.3.2 Approche par LLM et injection de lexique

Suite au bon résultat obtenu par le LLM, nous avons étudié la justification des réponses retournées par l'API (cf. Annexe A.2) et nous avons observé que le modèle allait puiser dans ses connaissances générales afin de lever les ambiguïtés mais semblait parfois incertain ou n'avait pas identifié dans son apprentissage la signification de l'acronyme. Nous avons alors émis l'hypothèse que l'injection d'un lexique défini par la SNCF elle-même serait bénéfique pour le modèle car il pourrait mieux guider sa prise de décision. Il existe une collection d'acronymes de la SNCF et leurs définitions qui est disponible en *open-data* (SNCF, 2017). Nous avons alors conçu l'approche suivante afin de l'incorporer. Plutôt que d'implémenter un système RAG dynamique qui consisterait à rechercher des documents pertinents à l'inférence, nous avons opté pour une **injection de lexique statique**.

Pour chaque acronyme du jeu de test, nous avons d'abord extrait toutes les définitions candidates correspondantes (l'acronyme et ses formes étendues) directement du fichier de lexique *open-data*. Ces candidats ont ensuite été injectés dans le **prompt** du LLM avec le contexte de l'acronyme. La tâche du modèle devenait alors fortement guidée par cette sélection de candidats. Pour cette approche, nous avons réalisé deux expériences principales :

— **Expérience 6 :** Combinaison du *prompt* de l'expérience 5 (CoT + exemples) avec l'injection de lexique améliorant le score à **84,25%**

— **Expérience 7 :** Suppression des exemples et CoT pour ne conserver que l'injection de lexique. De façon notable, cette approche obtient un F1-score de **85,21%**, améliorant de 1,45 point le score de l'expérience 5 et celui l'expérience 6 de 0.96 point, suggérant que pour cette tâche, fournir un ensemble délimité de faits (les candidats du lexique) était plus efficace que les exemples manuels de raisonnement.

4.3.3 Ajout d'une heuristique de sélection

En nous basant sur la découverte du biais de construction systématique identifié en Section 3.3 (21% des données de test), nous avons conçu une heuristique pour exploiter ce pattern. Pour résoudre l'ambiguïté, nous avons conservé l'**approche de l'expérience 7** (injection de lexique statique uniquement) en y ajoutant une heuristique de sélection post-traitement dans le cas où il existe un choix possible entre la version longue et courte afin de déterminer si l'une ou l'autre des versions était préférée dans les données de test (cf. Annexe A.3).

— **L'expérience 8**, appliquant une règle "du plus court", a présenté un F1-score affaibli de 63,16% (-22,05 points).

— **L'expérience 9**, appliquant la "règle du plus long", a obtenu un F1-score de 87,89% (+2,68 points).

L'expérience 9 obtient les meilleures performances (par rapport à nos expériences précédentes) en exploitant le biais de longueur et suggère que les données de référence privilégient systématiquement la définition la plus complète et explicative lorsque plusieurs options valides sont présentées. Afin de vérifier la stabilité de ce résultat, nous avons évalué sa robustesse par plusieurs itérations. Les résultats montrent un score moyen de 87,31% ± 0,38% (n=5) indiquant une légère variabilité malgré une température réglée à 0 pour l'inférence par le modèle.

Pour notre **expérience finale**, nous souhaitons évaluer une structuration plus explicite de l'instruction (*prompt*) passant d'une instruction directe à une décomposition du raisonnement par étapes afin de contraindre le modèle à une vérification de son choix par rapport au contexte initial en s'inspirant de la méthode (*Chain of Thought*) explorée en Section 4.3.1.

— **L'expérience 10**, reprenant l'expérience 9 en ajoutant une structuration explicite du raisonnement, a obtenu un F1-score maximal de **88,06%** (+0,17 point).

L'expérience finale (n°10) démontre que la structuration de la consigne de façon très explicite peut encore améliorer le score. La réplication de cette expérience afin de mesurer la robustesse de cet ajout donne un score moyen de 87,52% ± 0,56% (n=3). L'écart-type est légèrement supérieur, suggérant une sensibilité ou instabilité par rapport à l'expérience précédente. Cependant, le score moyen étant le meilleur que nous ayons obtenu à travers nos expériences,

nous conservons cette dernière comme notre meilleure soumission au classement public du défi. Nous obtiendrons un F1-score final de **88,14%** sur le classement privé [1].

5 Résultats et analyse

Cette section présente les performances quantitatives de nos expérimentations. Le tableau 1 résume les performances maximales (F-mesure sur le classement public) des principales approches décrites dans la section 4, organisées par ordre de complexité et de performance croissantes.

N° d'exp	Approche	Description détaillée	F1 public	F1 privé
1	Baseline	TF-IDF + Régression logistique	48,69%	51,65%
2	Fine-tuning (Gold)	CamemBERT-base (sur données "gold")	49,33%	53,00%
4	Fine-tuning (Silver)	ModernCamemBERT (Gold + Silver data, k-fold)	79,21%	78,22%
5	LLM (CoT)	Gemini 2.5 Pro (CoT + exemples manuels)	83,76%	87,75%
6	LLM (CoT + lexique)	Exp. 5 + Injection de lexique statique	84,25%	**90,08%**
7	LLM (Lexique seul)	Injection de lexique statique sans exemples manuels	85,21%	88,54%
9	LLM (Lexique + heuristique)	Exp. 7 + Règle de sélection "du plus long"	87,89%	87,79%
10	LLM (Lexique + heuristique)	Exp. 9 + Structure raisonnement explicite	**88,06%**	88,14%

TAB. 1 – *Résumé des performances des expériences.*

L'analyse de ces résultats montre une progression nette. Le passage de la *baseline* (Exp. 1) à une approche *fine-tunée* sur données augmentées (Exp. 4) apporte le gain le plus substantiel (+30,52 points). L'utilisation d'un LLM (Exp. 5) surpasse d'emblée cette approche (+4,55 points). Par ailleurs, l'utilisation de connaissances factuelles couplée à une heuristique issue de l'observation des données (Exp. 9) s'avère plus efficace que la combinaison des connaissances factuelles avec les exemples de raisonnement fournis dans l'expérience 6 avec un gain de 3,64 points. Si l'expérience 6 enregistre notre pic de performance absolu sur le classement privé (90,08%), elle présente un écart significatif (+5,83 points) par rapport au classement public suggérant une instabilité ou une sensibilité favorable à la distribution spécifique à la partie privée du jeu d'évaluation. À l'inverse, nous désignons l'Expérience 10 comme notre meilleure approche globale, atteignant notre meilleur score sur le classement public (88,06%), avec une capacité de généralisation supérieure indiquée par l'écart le plus faible de toutes nos expériences (+0,08 point) entre les deux parties du jeu d'évaluation.

6 Discussion et limitations

Les expérimentations menées ont démontré que les meilleurs résultats ont été obtenus avec l'expérience 10 (Section 4.3.3) qui nous place en deuxième position de la compétition. L'analyse des résultats du tableau 1 révèle plusieurs points clés.

1. Il est intéressant de noter qu'après la clôture officielle du défi, nous avons relancé strictement le même protocole (Expérience 10) en utilisant une itération plus récente du modèle : Gemini 3 (Pro). Sans aucune modification du *prompt*, le score est monté à 89,19% sur le classement public et **91,24%** sur le classement privé (améliorant nos précédents scores de +1,30 points et +3,10 points respectivement). Ce résultat souligne la forte corrélation entre les résultats obtenus et la capacité intrinsèque du modèle à suivre des instructions conditionnelles fines, capacité qui progresse rapidement avec les nouvelles versions.

Premièrement, la faible augmentation de score entre la *baseline* et le *fine-tuning* de CamemBERT-base est cohérente avec la faible quantité de données à disposition et de la disparité entre le jeu d'entraînement et de test relevée en section 3.

Deuxièmement, la génération de données "silver" (Exp. 4) a permis de franchir un premier palier de performance significatif. Cela démontre que même des données auto-annotées, si elles sont de qualité suffisante, peuvent pallier la rareté des annotations manuelles pour l'ajustement fin d'un modèle pré-entrainé.

Troisièmement, l'utilisation directe d'un LLM (Gemini 2.5 Pro) en explicitant le raisonnement (Exp. 5) a permis d'obtenir de meilleurs résultats que l'approche par *fine-tuning* soulignant la capacité de raisonnement en contexte de ces modèles à grande échelle, y compris pour cette tâche.

Cependant, il est intéressant de noter l'intérêt de l'injection de connaissances externes. En émettant l'hypothèse que la ressource *open-data* de la SNCF (SNCF, 2017) était suffisamment à jour et détaillée pour définir les acronymes du défi, nous avons testé une approche par injection de lexique statique. Le fait que l'expérience 7, utilisant uniquement ce lexique sans exemples issus des données gold, surpasse l'expérience 6 qui combine exemples (*few-shots* et raisonnement) et lexique suggère que la présence des faits (le lexique) est plus déterminante que les exemples de raisonnement pour ce modèle. Cette injection de lexique sera conservée jusque dans la meilleure expérience présentée.

Enfin, l'approche finale présente des limites inhérentes :

— **Coût et vitesse :** L'inférence via l'API d'un LLM avec un prompt contenant un extrait ciblé d'un lexique statique est coûteuse et lente ce qui peut être un frein pour une application industrielle à grande échelle comparé à un modèle ajusté local.

— **Spécificité :** Notre meilleure approche dépend fortement de l'existence d'un lexique *open-data* correspondant aux données du défi. Son efficacité n'est pas garantie sur des acronymes propriétaires absents de ce lexique.

— **Reproductibilité :** Même en ayant fixé la température du modèle à 0, une variabilité stochastique minime (mais non nulle) a été observée dans les scores (section 4.3.3). C'est un facteur important à considérer dans l'objectif d'une mise en production qui reposerait sur des appels à l'API du modèle fermé Gemini.

7 Conclusion et perspectives

Cet article présente la démarche itérative adoptée pour le défi TextMine EGC 2026 (Lefeuvre et al., 2025) visant à résoudre la désambiguïsation d'acronymes en domaine ferroviaire. En partant d'une *baseline* TF-IDF, nous avons exploré l'ajustement fin de modèles pré-entrainés avant de nous tourner vers les LLM. Notre système final atteint une **F-mesure maximale de 88,14%** sur le classement privé. Il repose sur l'utilisation du modèle Gemini 2.5 Pro, auquel nous fournissons en contexte un lexique statique issu des données *open-data* de la SNCF combiné à une heuristique de sélection de la forme étendue la plus longue et une structuration explicite du raisonnement à suivre. Cette étude suggère que pour cette tâche spécifique, l'injection directe de connaissances factuelles pertinentes dans le prompt s'avère plus efficace que des techniques de raisonnement complexes (CoT) ou des approches supervisées limitées par la quantité et diversité des données disponibles lors de ce défi. Plusieurs pistes

sont envisageables pour les travaux futurs. La plus prometteuse consisterait en la **distillation de connaissances** : employer notre modèle le plus performant (Exp. 10) pour générer des labels de haute qualité pour l'ensemble du jeu d'entraînement. Ces données pourraient alors être utilisées pour ajuster finement un modèle plus petit (comme ModernCamemBERT) combinant ainsi la précision du LLM "ancré" au lexique avec l'efficacité d'un modèle supervisé compact, à notre sens plus adapté à une mise en production.

8 Disponibilité des ressources

Une sélection pertinente des *scripts*, *prompts* et données est accessiblement publiquement sur le dépôt suivant : `https://github.com/Emvista/TextMine-EGC-2026/`

9 Bibliographie

Références

Antoun, W., B. Sagot, et D. Seddah (2025). ModernBERT or DeBERTaV3 ? Examining Architecture and Data Influence on Transformer Encoder Models Performance. arXiv :2504.08716 [cs].

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, et D. Amodei (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Volume 33, pp. 1877–1901. Curran Associates, Inc.

Ciosici, M., T. Sommer, et I. Assent (2019). Unsupervised Abbreviation Disambiguation Contextual disambiguation using word embeddings. arXiv :1904.00929 [cs].

David, R., A. Kernerman, I. Kernerman, N. Ferranti, et A. Siani (2024). Multilingual Word Sense Disambiguation for Semantic Annotations : Fusing Knowledge Graphs, Lexical Resources, and Large Language Models.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv :1810.04805 [cs].

Ding, Y., W. Fan, L. Ning, S. Wang, H. Li, D. Yin, T.-S. Chua, et Q. Li (2024). A Survey on RAG Meets LLMs : Towards Retrieval-Augmented Large Language Models. arXiv :2405.06211 [cs] version : 1.

Gemini Team (2025). Gemini 2.5 : Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv :2507.06261 [cs] version : 1.

Kugic, A., S. Schulz, et M. Kreuzthaler (2024). Disambiguation of acronyms in clinical narratives with large language models. *Journal of the American Medical Informatics Association : JAMIA 31*(9), 2040–2046, doi: 10.1093/jamia/ocae157.

Lackner, A., A. Vega-Wilson, et C. Lang (2025). Terminology Augmented Generation : A Systematic Review of Terminology Formats for In-Context Learning in LLMs.

Lefeuvre, L., C. Reutenauer, A. Guille, P. Cuxac, et C. Lopez (2025). Défi TextMine / EGC 2026.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, et D. Kiela (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Volume 33, pp. 9459–9474. Curran Associates, Inc.

Li, Y., B. Zhao, A. Fuxman, et F. Tao (2018). Guess Me if You Can : Acronym Disambiguation for Enterprises. In I. Gurevych et Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Melbourne, Australia, pp. 1308–1317. Association for Computational Linguistics, doi: 10.18653/v1/P18-1121.

Liu, S. et Y. Fang (2024). Use Large Language Models for Named Entity Disambiguation in Academic Knowledge Graphs. In G. Guan, C. Kahl, B. Majoul, et D. Mishra (Eds.), *Proceedings of the 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023)*, Volume 16, pp. 681–691. Dordrecht : Atlantis Press International BV, doi: 10.2991/978-94-6463-264-4$_7$9. Series Title : Atlantis Highlights in Computer Sciences.

Martin, L., B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, V. d. l. Clergerie, D. Seddah, et B. Sagot (2020). CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7203–7219, doi: 10.18653/v1/2020.acl-main.645. arXiv :1911.03894 [cs].

Moon, S., S. Pakhomov, et G. B. Melton (2012). Automated Disambiguation of Acronyms and Abbreviations in Clinical Texts : Window and Training Size Considerations. *AMIA Annual Symposium Proceedings 2012*, 1310–1319.

Pan, C., B. Song, S. Wang, et Z. Luo (2021). BERT-based Acronym Disambiguation with Multiple Training Strategies.

Siriwardhana, S., R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, et S. Nanayakkara (2023). Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Transactions of the Association for Computational Linguistics 11*, 1–17, doi: 10.1162/tacl$_{a}$0530. Place : Cambridge, MA Publisher : MIT Press.

SNCF (2017). Lexique des abréviations SNCF.

Vogel, J. (2025). Leveraging Large Language Models for Generative Acronym Disambiguation in the Biomedical Domain.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, et D. Zhou (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv :2201.11903 [cs].

A Exemples de données, justifications et prompts

A.1 Exemple d'exemple pour le *few-shot*

Le bloc ci-dessous montre un exemple utilisé pour l'apprentissage *few-shot* (Expérience 5, Section 4.3.1) illustrant un cas où le contexte est très clair.

```
{
    "acronym": "AC",
    "text": "1.3. Résumé des modifications Suppression des
        BB75000     75100     75400 suite à la parution
        d une  AC nationale. Restrictions EURO4000-II.  RA
       -RT-5151B- Version 02 du 03-04-2017 Page 1",
    "options": {
        "Amélioration Continue": False,
        "Autorité de Certification": False,
        "Agent d'aCcompagnement ": False,
        "AURILLAC": False,
        "Attestation de Compatibilité du matériel roulant
           à l'infrastructure": True,
        "Accord Cadre": False,
        "Agent du service Commercial ": False,
        "Agent Circulation ": False,
        "Action Corrective": False,
        "ACcès": False
    },
    "raisonnement": "Le contexte mentionne la parution d'
       une 'AC' qui indique des modifications à ce qui
       semble être des types de trains. Seule la
       proposition 'Attestation de Compatibilité du maté
       riel roulant à l'infrastructure' indique un
       document susceptible de contenir des modifications,
        dont des suppressions."
}
```

A.2 Exemple de justification CoT (sortie brute)

Extrait de la sortie JSON brute du modèle (Expérience 5) incluant le résumé de la réflexion (*Chain-of-Thought*) générée.

```
{"response":{"candidates":[{"content":{"parts":[{"text":

"Réflexion: Le contexte mentionne une restriction "en UM2 et
   UM3", ce qui dans le jargon ferroviaire désigne l'
   accouplement de plusieurs rames (ou éléments) pour former
   un seul train, rendant "Unité-Multiple de trois éléments"
   la définition la plus appropriée.
```

```
```json{"index_correct": 1}```"}],

"role":"model"},"index":0,"finishReason":"STOP"}],"
    usageMetadata":{"promptTokenCount":1245,"totalTokenCount
    ":2404,"thoughtsTokenCount":1077,"candidatesTokenCount
    ":82,"promptTokensDetails":[{"modality":"TEXT","tokenCount
    ":1245}]},"responseId":"3J0EadKGF6mwqtsP35-X6QQ","
    modelVersion":"gemini-2.5-pro"},"custom_id":280}
```


A.3 *Prompt* de l'expérience 9

Prompt complet utilisé pour l'expérience 9 (LLM avec injection de lexique statique + heuristique).
Tu es un expert en désambiguïsation d'acronymes dans des documents techniques ferroviaires. Ta mission est de choisir la bonne définition d'un acronyme parmi une liste, en te basant sur deux sources : le contexte du texte et un lexique de référence.

Ta méthode est la suivante :

1. **Contexte 1 (Texte) :** Analyse le texte où l'acronyme apparaît.
2. **Contexte 2 (Lexique) :** Consulte les définitions fournies ci-dessous, issues d'un lexique SNCF. Elles sont une aide pour confirmer ton choix ou résoudre un doute. Le contexte du texte prime si le lexique est ambigu ou contradictoire.
3. **RÈGLE IMPORTANTE :** Si plusieurs options te semblent correctes et décrivent la même entité (par exemple, une version courte « Réseau Ferré » et une version longue « Réseau Ferré. Description... »), **tu dois TOUJOURS privilégier l'option la plus longue et la plus descriptive.**
4. **Réflexion :** Analyse brièvement ton choix en mentionnant la Règle 3 si tu l'as appliquée.
5. **Sortie :** Fournis ta réponse finale sous la forme d'un objet JSON valide contenant uniquement la clé « index_correct ». La valeur est l'index (un entier) de l'option choisie.

S'il n'y a pas de bonne réponse, réponds avec l'index -1.

EXEMPLE À TRAITER
Contexte 1 (Texte) :

`{example_json}`

Contexte 2 (Lexique SNCF pour l'acronyme « `{acronym}` ») : `{lexicon_context}`
Sortie attendue :

Summary

The TextMine EGC 2026 Challenge, proposed by SNCF, aims to address the complex problem of disambiguating polysemous acronyms within regulatory texts in the railway domain. This task, which lies at the intersection of named entity recognition and entity linking,
```

is of crucial importance for information extraction in industrial settings. Its resolution requires a fine-grained understanding of context in order to correctly associate an acronym with its expanded form, potentially selected from a list of provided candidates. This paper presents a comparative study of the methodologies implemented by the author ("Mokipo_" team) to tackle this challenge, particularly by exploring the capabilities of a large language model (LLM) on this task. The results obtained (F1-score of 88,14%, second place on the private leaderboard) demonstrate that, for this specialized disambiguation task, the contextual reasoning ability of LLMs—activated through advanced prompting techniques—significantly outperforms the supervised classification approaches tested.