Annotation linguistique sous incertitude structurelle : vers un modèle ensembliste basé sur les fonctions de croyance

Anna Pappa[1,2], Pierre Pomeret-Coquot[3], Francis Faux[2], Revekka Kyriakoglou[1]

[1] LIASD, Université Paris 8, 2 rue de la liberté, 93526 Saint Denis, France
[2] IRIT, Université de Toulouse, CNRS, Toulouse INP, Toulouse, France
[3] IRIT, Université de Toulouse, CNRS, Toulouse INP, UT Capitole, IUT de Rodez, Toulouse, France

ap@up8.edu – francis.faux@irit.fr – pierre.pomeret@irit.fr – kyriakoglou@up8.edu

Résumé

La majorité des travaux en annotation linguistique modélisent l'assignation d'une étiquette comme une application déterministe associant chaque instance à une unique catégorie. Cette hypothèse suppose que les catégories sont indépendantes les unes des autres et qu'une même unité ne peut appartenir qu'à une seule catégorie.

À partir d'un cas d'étude portant sur l'annotation fonctionnelle de prompts produits en contexte d'apprentissage, nous montrons que dans certains schémas catégoriels, l'indétermination est structurelle et provient du recouvrement partiel des catégories.

Nous étudions comment la théorie des fonctions de croyance permet de décrire avec exactitude les différents ensembles d'étiquettes proposés par les annotateurs, en représentant explicitement l'incertitude et les désaccords.

Cette annotation ensembliste, bien que non univoque, conserve l'information interprétative et permet d'extraire plusieurs sous-corpus annotés et d'en mesurer la qualité. Nous identifions également des regroupements pertinents de catégories d'étiquettes, permettant de réduire l'incertitude sans perdre en granularité sémantique.

Le corpus annoté obtenu est donc exact et indépendant des tâches d'apprentissage auxquelles il est destiné, tout en offrant un cadre susceptible d'améliorer la précision des modèles et d'affiner leur évaluation.

Mots-clés

Annotation TALN Incertitude Fonctions de croyance

Abstract

Most works in linguistic annotation models label assignment as a deterministic function that associates each instance to a single category. This assumption implies that categories are independent from one another and that a given annotation unit belongs to only one label.

Based on a case study involving the functional annotation of prompts produced in a learning context, we show that in certain categorial schemes, indeterminacy is structural and arises from partial overlap between categories. We study how the theory of belief functions can accurately describe the different sets of labels proposed by annotators, by explicitly representing uncertainty and disagreement.

Although non-univocal, this set-valued annotation preserves interpretative information and allows us to extract several annotated subcorpora and measure their quality. We also identify relevant groupings of labels that reduce uncertainty without losing semantic granularity.

The resulting annotated corpus is therefore accurate and independent of the learning tasks for which it is intended, while providing a framework that can improve model accuracy and refine model evaluation.

Keywords

Annotation NLP Uncertainty Belief functions

1 Introduction

Ce travail se situe à l'intersection de la théorie de l'annotation en TALN (méthodologique), de la modélisation formelle de l'incertitude (incertitude épistémique structurelle) et de l'étude des mécanismes décisionnels humains (décisions d'annotation) dans les processus de catégorisation (processus de décision sous indétermination interprétative). De nombreux travaux en annotation linguistique et fonctionnelle reposent sur l'hypothèse implicite d'une correspondance déterministe (univoque) entre une instance et une catégorie [4, 23]. Or, dans plusieurs contextes d'annotation fonctionnelle, les prédicats définissant les labels ne sont pas strictement disjoints : une même instance peut satisfaire simultanément plusieurs critères catégoriels valides.
Formellement, l'annotation est modélisée comme une fonction

$$f : I \to L,$$

où I désigne l'ensemble des instances annotées et L l'ensemble des étiquettes du schéma d'annotation [14]. Dans un contexte d'apprentissage supervisé, ces étiquettes correspondent aux labels associés aux instances du jeu de données [17]. Dans cet article, nous utilisons le terme *label* pour désigner les catégories du schéma d'annotation, afin de conserver une cohérence avec leur utilisation ultérieure comme variables cibles dans les jeux de données d'apprentissage supervisé.

Nous examinons un cas d'usage concret : l'annotation fonctionnelle d'interactions (ou prompts) avec des IA génératives, produites par des étudiants lors d'un hackathon. Le schéma d'annotation, inspiré de [13], repose sur 21 catégories fonctionnelles et une assignation mono-label.

Nous montrons que, malgré un protocole rigoureux, certaines interactions ne peuvent être assignées de manière univoque à un seul label.

Nous soutenons que cette situation ne relève pas nécessairement d'une erreur d'annotation ni d'un manque d'information, mais d'une propriété structurelle de l'espace catégoriel. L'annotation doit alors être modélisée non comme un problème de classification déterministe, mais comme un processus de décision sous incertitude interprétative, correspondant à une attribution d'incertitude structurée.

Plutôt que de modéliser l'assignation comme une fonction à valeur unique, nous proposons une fonction ensembliste

$$F : I \to \mathcal{P}(L),$$

où $\mathcal{P}(L)$ désigne l'ensemble des parties de L. L'annotation devient alors une relation entre interactions et sous-ensembles de labels compatibles.

Nous introduisons également, pour chaque instance annotée, un score dans l'intervalle $[0, 1]$, correspondant à la confiance de l'annotateur dans sa compréhension du prompt. Ce score ne porte donc pas sur la validité d'un label donné, mais sur l'intelligibilité de l'instance elle-même. Il permet de distinguer deux formes d'incertitude : d'une part, l'incertitude interprétative liée à une compréhension partielle ou fragile du prompt ; d'autre part, l'indétermination catégorielle, qui subsiste même lorsque le prompt est bien compris, du fait du recouvrement entre labels.

Cette reformulation permet de représenter explicitement la pluralité des labels valides pour une même instance. Nous introduisons la notion de *complexité interprétative*, définie pour une instance i par :

$$C(i) = |F(i)|\,,$$

afin de caractériser le degré d'indétermination catégorielle. Ce cadre formel est opérationnalisé au moyen des fonctions de croyance issues de la théorie de Dempster-Shafer [8, 24], qui permettent de représenter et de combiner cette incertitude sur les ensembles de labels.

1.1 Travaux relatifs

Trois axes de recherche sont directement pertinents pour notre approche : les limites des schémas d'annotation mono-label, la reconsidération du désaccord d'annotation, et les conséquences pour la modélisation en apprentissage automatique.

1.1.1 Annotation mono-label et ses limites

L'hypothèse d'un label unique par instance suppose implicitement que les prédicats définissant les catégories sont mutuellement exclusifs et exhaustifs. Or cette hypothèse ne tient pas dans tous les contextes.

Si la multifonctionnalité des énoncés a été reconnue théoriquement dès les premières formalisations des actes de dialogue [1, 5] et formalisée dans le standard ISO [6], la pratique d'annotation dominante a néanmoins maintenu l'assignation mono-label [4, 23]. Cette persistance s'explique par des contraintes opérationnelles, telles que le coût de l'annotation et la complexité de l'évaluation inter-annotateurs, plutôt que par une adhésion théorique à l'hypothèse déterministe.

1.1.2 "Désaccord annotatif" comme signal structurel

Dans des travaux de recherche récents, on trouve un courant qui remet en question l'interprétation du désaccord annotatif comme une erreur à éliminer [2, 27]. [2] identifie plusieurs mythes fondateurs de la pratique d'annotation, dont celui du *gold standard* unique, et proposent de traiter le désaccord comme une information à modéliser. [19] montre empiriquement que le désaccord sur des tâches d'inférence textuelle reflète des divergences interprétatives légitimes plutôt que des erreurs. [20] généralise ce constat à l'ensemble des pratiques d'annotation en NLP et plaide pour une révision profonde de la notion de vérité terrain.

Ces travaux partagent l'idée que le désaccord est informatif. Ils diffèrent de notre approche sur un point essentiel : ils traitent le désaccord comme un phénomène *inter-annotateurs*, lié à la variation des jugements humains.

Nous proposons ici une lecture complémentaire : l'indétermination peut être une propriété *intra-instance*, indépendante du nombre d'annotateurs ou de leurs divergences, et inhérente au recouvrement partiel des catégories du schéma d'annotation lui-même. Dans ce sens, le caractère *structurel* du désaccord ne renvoie pas à un manque d'information ni à une erreur d'annotation, mais à la structure de l'espace catégoriel.

1.1.3 Conséquences pour la modélisation en apprentissage automatique

Lorsque plusieurs labels peuvent être compatibles avec une même instance, le problème ne relève plus de la classification mono-label classique. La littérature en apprentissage automatique traite ces situations dans le cadre de la *multi-label classification*, où une instance peut être simultanément associée à plusieurs labels [26]. Dans la plupart des travaux existants, cette pluralité de labels est mise en évidence empiriquement à partir des données annotées, par exemple lorsque plusieurs annotateurs produisent des jugements divergents qui sont conservés sous forme de distributions ou d'ensembles de labels [2, 27]. Dans ce cas, la pluralité est interprétée comme résultante de la variation inter-annotateurs. Dans le contexte de certaines tâches d'annotation linguistique, plusieurs travaux ont également montré que le choix forcé entre catégories peut masquer une ambiguïté réelle et introduire un biais dans l'évaluation de la fiabilité annotative [21]. L'approche proposée ici repose sur une hypothèse différente : la pluralité des labels ne résulte pas d'un agrégat de jugements d'annotation mais d'une propriété structurelle de l'espace catégoriel.

1.2 Positionnement par rapport aux pratiques existantes

Dans la pratique courante de l'annotation fonctionnelle, le désaccord d'annotation est généralement interprété comme une erreur ou une divergence à résoudre, et évalué au moyen de mesures de fiabilité telles que le coefficient κ [7, 3]. Cette hypothèse d'un "gold label" unique fait cependant l'objet de discussions croissantes [2, 27, 20] : le désaccord entre annotateurs y est reconsidéré comme un signal potentiellement informatif plutôt que comme un bruit à éliminer. Ces travaux expliquent toutefois l'indétermination principalement par la variation des jugements humains. Dans ce cadre, l'objectif est de réduire l'indétermination.

Nous proposons ici une lecture complémentaire : l'indétermination peut également être une propriété *intra-instance*, révélant un recouvrement partiel des catégories du schéma d'annotation lui-même. Dans ce cas, la pluralité des labels ne doit pas être réduite à une décision exclusive mais peut être explicitement représentée et intégrée au processus annotatif. Cette observation conduit à reformuler l'annotation comme une relation entre instances et labels. Pour représenter et opérationnaliser cette incertitude structurelle, nous recourons à la théorie des fonctions de croyance [8, 24], qui offre un cadre formel permettant d'attribuer des degrés de confiance à des ensembles de labels plutôt qu'à des labels uniques.

2 Cadre d'annotation existant

2.1 Schéma d'annotation

L'annotation (type de catégorie) des interactions s'inspire du protocole proposé par Guner et al. [13] développé pour l'analyse des échanges étudiants-outils IA générative (ChatGPT, Gemini, etc.) dans un contexte d'apprentissage de la programmation. Dans ce schéma, chaque prompt constitue l'unité d'annotation. Les auteurs définissent 21 catégories fonctionnelles, visant à caractériser la *finalité* du prompt dans l'interaction. L'annotation repose sur une analyse pragmatique du contenu de la requête, centrée sur l'intention communicative exprimée par l'étudiant, indépendamment des résultats obtenus ou de la performance académique ultérieure. Les catégories couvrent un ensemble de fonctions interactionnelles comme la reproduction des instructions, la demande d'explication conceptuelle, la révision de code, la définition du contexte, etc. (cf. Table 1). Dans le protocole original, chaque prompt est assigné à une unique catégorie fonctionnelle.

L'opération de labellisation est réalisée par deux experts, puis discutée jusqu'à l'obtention d'un consensus. Cette procédure vise à garantir la cohérence interprétative et à limiter les biais individuels.

2.2 Adaptation au cas du hackathon

Dans notre étude, ce schéma d'annotation est appliqué à un corpus de prompts produits lors d'un hackathon étudiant dans le cadre d'un cours d'IA à l'université. L'annotation initiale repose sur une assignation mono-label : chaque prompt est annoté indépendamment par deux experts, puis les divergences éventuelles sont arbitrées par un troisième annotateur afin d'aboutir à un label unique.

Toutefois, au cours du processus d'annotation, nous avons observé plusieurs cas où plusieurs labels apparaissaient simultanément compatibles avec une même instance. Ces situations ne résultaient pas d'un manque d'information ni d'un désaccord interprétatif entre les annotateurs, mais de la compatibilité sémantique entre plusieurs catégories du schéma.

Afin d'explorer cette indétermination, nous avons également réalisé une seconde annotation autorisant l'assignation de plusieurs labels par prompt, accompagnée d'un degré de confiance pour chaque attribution.

Dans ce travail, l'analyse porte exclusivement sur le contenu des prompts eux-mêmes : les interactions sont considérées indépendamment de leur contexte conversationnel ou du profil des étudiants.

3 Indétermination structurelle

Dans ce travail, nous considérons que l'annotation d'une interaction $i \in I$ peut être associée non pas à un label unique, mais à un ensemble de labels plausibles $F(i) \subseteq L$, tel qu'introduit précédemment. Cette représentation permet de rendre explicites les situations dans lesquelles plusieurs catégories apparaissent simultanément compatibles avec une même instance.

Même avec une guideline claire, une consigne linguistique stricte et l'expertise des annotateurs, certaines interactions peuvent être interprétées de plusieurs façons fonctionnelles valides.

3.1 Non-orthogonalité des catégories

Les catégories définies dans le schéma d'annotation couvrent un ensemble de fonctions interactionnelles relativement fines. Cependant, certaines d'entre elles ne sont pas mutuellement exclusives sur le plan interprétatif.

Un même prompt peut ainsi relever simultanément de plusieurs catégories fonctionnelles. Par exemple, une interaction peut être à la fois une *Coding task inquiry*, une *Concept explanation* et une *Defining context*. Cette superposition résulte du fait que l'espace catégoriel n'est pas orthogonal : les prédicats définissant les labels se recouvrent partiellement. L'indétermination observée ne provient pas nécessairement d'une divergence entre annotateurs, mais de la structure du schéma d'annotation.

3.2 Complexité interprétative

Dans ce cadre, il devient possible de caractériser le degré d'indétermination associé à une interaction.

Prenons l'exemple :

> "Is this correct ?"

Cet énoncé peut correspondre à une vérification, à une demande implicite d'explication ou à une validation finale. Nous définissons la complexité interprétative d'une interaction i comme :

$$C(i) = |F(i)|$$

Type de prompt	Explication de prompt
1. Exact copy of instruction	Directly copy-pasting entire task instructions.
2. Partial copy of instruction	Partially copy-pasting task instructions.
3. Summary of instruction	Summarizing or paraphrasing task instructions.
4. Coding task inquiry	Asking for specific coding knowledge or guidance related to a particular task or requirement.
5. Refining GPT code	Asking for refinement on the GPT solution to meet the task requirements.
6. Refining own code	Asking for refinement on their own code to meet the task requirements.
7. Debugging own code	Asking for finding errors in their own code.
8. Debugging GPT code	Asking for fixing errors in GPT solution.
9. Reviewing own code	Requesting review, enhancement, or optimization for their own code.
10. Reviewing the GPT code	Requesting review, enhancement, or optimization for GPT code.
11. GPT code explanation	Asking for explanation about GPT solution.
12. Prompt revision	Requesting GPT to modify its previous response based on feedback or clarification.
13. Concept explanation	Asking for explanation on related concepts.
14. Asking for more info	Asking for clarification or additional information about a given response.
15. Generating alternative code	Asking GPT to simplify its previous solution or to offer another alternative solution.
16. Defining context	Establishing the context, scope, or requirements before the main prompt.
17. Combining all GPT code	Asking to combine previous GPT responses into one.
18. Code for given output	Asking GPT to code for a given output.
19. Verifying GPT solution	Asking GPT for verification of its solution.
20. Suggested inquiry	Suggested prompts from the guide document.
21. Other	Typos, spelling errors, obscure statements, greetings, or gratitude expressions.

TABLE 1 – Catégories fonctionnelles (des prompts) et leurs définitions ([13]).

où $F(i)$ désigne l'ensemble des labels plausibles. Lorsque $C(i) = 1$, l'assignation est univoque. Lorsque $C(i) > 1$, l'interaction présente une indétermination structurelle que le schéma mono-label force artificiellement à réduire. Cette pluralité d'interprétations ne relève pas d'une ambiguïté lexicale mais de la coexistence de plusieurs dimensions fonctionnelles au sein d'un même énoncé et doit pouvoir être représentée explicitement dans le modèle d'annotation.

4 Formalisation dans le cadre des fonctions de croyance

4.1 Théorie des fonctions de croyance

L'interprétation "probale" (*evidential*) de la théorie des fonctions de croyance [8, 24] permet de modéliser les degrés de confiance issus des annotations ensemblistes effectuées par les experts. Introduisons d'abord quelques définitions. Soit $i \in I$ une interaction. Le *cadre de discernement* est l'ensemble des labels $L = \{1, \ldots, n\}$ pouvant décrire i. Une fonction de masse $m_i : \mathcal{P}(L) \to [0, 1]$ vérifie $m_i(\emptyset) = 0$ et $\sum_{B \subseteq L} m_i(B) = 1$. Chaque ensemble $B \subseteq L$ tel que $m_i(B) > 0$ est appelé un "élément focal de m_i" et représente le fait que l'on dispose de l'information $F_i = B$ avec un degré de confiance égal à $m(B) \in [0, 1]$. Intuitivement, B est l'ensemble de labels donné par un expert.

À partir de la fonction de masse m_i, on peut définir deux mesures, Bel_i et $\mathrm{Pl}_i : \mathcal{P}(L) \to [0, 1]$, appelées respectivement fonctions de croyance et de plausibilité, définies par :

$$\mathrm{Bel}_i(A) = \sum_{B \subseteq A} m_i(B) \quad \text{et} \quad \mathrm{Pl}_i(A) = \sum_{B \cap A \neq \emptyset} m_i(B)$$

Ces mesures sont duales : $\mathrm{Bel}_i(A) + \mathrm{Pl}_i(\bar{A}) = 1$ pour tout $A \subseteq L$. $\mathrm{Bel}_i(A)$ indique à quel point l'ensemble d'étiquettes $A \subseteq L$ est impliqué par les annotations des experts, et $\mathrm{Pl}_i(A)$ indique à quel point il est compatible (cf. exemple 1). Notons que nous avons deux familles de cas particuliers : si tous les experts font des annotations précises, alors les éléments focaux sont des singletons ($\forall B$ focal, $|B| = 1$) et $\mathrm{Bel}_i = \mathrm{Pl}_i$ est une mesure de probabilité sur L ; par ailleurs si tous les experts font des annotations compatibles entre elles, alors les éléments focaux sont emboîtés ($\forall B_1$ et B_2 focaux, $B_1 \subseteq B_2$ ou $B_2 \subseteq B_1$) et Pl_i est une mesure de possibilité et Bel_i est sa mesure de nécessité duale [12].

Example 1 (Annotation) *Considérons l'annotation de l'interaction i, avec $L = \{1, 2, 3\}$.*

— *L'expert a dit que $F_a(i) = \{1, 2\}$.*

— *L'expert b dit que $F_b(i) = \{2, 3\}$.*

On accorde autant de confiance à chaque expert, ce qu'on modélise par la fonction de masse m_i telle que :

$$m_i(\{1, 2\}) = m_i(\{2, 3\}) = 0.5 \qquad m_i(B) = 0 \text{ sinon.}$$

Comme on l'observe dans la table 2, la fonction de croyance (resp. de plausibilité) issue de m_i indique le degré d'implication (resp. de compatibilité) avec les annotations effectuées par les experts, pour chaque label et ensemble de labels. Ainsi, par exemple, le label 2 est complètement compatible ($\mathrm{Pl}_i(\{2\}) = 1$). L'ensemble de labels $\{1, 2\}$ est impliqué à 50% ($\mathrm{Bel}(\{1, 2\} = 0.5$) car impliqué par l'expert a uniquement. Notons que $\emptyset$ est nécessairement incompatible (contradiction), et que L est nécessairement certain (tautologie).

A	$m_i(A)$	$\mathrm{Bel}_i(A)$	$\mathrm{Pl}_i(A)$
$\emptyset$		0	0
$\{1\}$		0	0.5
$\{2\}$		0	1
$\{3\}$		0	0.5
$\{1,2\}$	0.5	0.5	1
$\{1,3\}$		0	1
$\{2,3\}$	0.5	0.5	1
$\{1,2,3\}$		1	1

TABLE 2 – Masses et degrés de croyance et de plausibilité pour l'exemple 1

Interprétation des valeurs. Dans l'interprétation probale, on suppose habituellement qu'un état (ici, une étiquette) $l^* \in L$ est le « vrai », que tout ensemble $A \subseteq L$ décrit la proposition « $l^* \in A$ » et que les valeurs $\mathrm{Bel}(A)$ et $\mathrm{Pl}(A)$ décrivent la croyance en cette proposition. Or, ici, nous nous intéressons à l'incertitude provenant de l'indétermination structurelle, et la « vérité » est un ensemble $A^* \subseteq L$ contenant toutes les étiquettes correctes. Ainsi, tout ensemble A représente la proposition « A contient l'une des étiquettes correctes » (i.e., $A \cap A^* \neq \emptyset$), et $\mathrm{Bel}(A)$ (resp. $\mathrm{Pl}(A)$) indique à quel point cette proposition est impliquée (resp. compatible) avec la connaissance.
Notons que cette interprétation diffère de l'interprétation en termes de famille de probabilités (ou *ensemble crédal*). Dans l'exemple 1, $\mathrm{Bel}(\{1,2\}) = 0.5$ et $\mathrm{Pl}(\{1,2\}) = 1$ indiquent que la moitié des annotations implique les étiquettes $\{1,2\}$ et que la totalité des annotations sont compatible avec ces étiquettes. Cependant, cela ne signifie généralement pas que la probabilité que l'étiquette correcte soit 1 ou 2 est comprise entre 0.5 et 1. En effet, l'intérêt du cadre proposé est de gérer l'incertitude structurelle, où une interaction ne possède pas forcément d'étiquette unique : une interaction peut avoir plusieurs étiquettes valides simultanément (par exemple, {1,2} ou {2,3}), ce qui n'est pas représentable par une distribution de probabilité.

4.2 Protocole d'annotation

On définit un protocole d'annotation (Algorithme 1) dans lequel l'expert peut attribuer à chaque instance soit un label unique, soit un ensemble de labels plausibles. À chaque annotation est associé un degré de confiance $d_i^x \in [0,1]$, reflétant le niveau de confiance de l'expert x dans la compréhension du prompt i, c'est-à-dire la clarté et la non-ambiguïté de l'énoncé. Le degré de confiance dans la compréhension du prompt est indépendant du nombre de labels attribués. Autrement dit, l'annotateur peut être très sûr de sa compréhension (degré proche de 1) tout en considérant que plusieurs catégories sont simultanément pertinentes.
Il serait également possible de donner un degré de confiance pour chacun des labels et/ou sous-ensembles des labels choisis, afin d'exprimer plus finement la certitude de l'expert. Toutefois, l'introduction de cette dimension compliquerait la définition du degré de confiance, rendrait la tâche d'annotation fastidieuse et entraînerait une plus grande complexité ; nous choisissons donc de ne pas la prendre en compte dans ce travail : chaque annotateur donne un unique ensemble d'étiquettes et un unique degré de confiance.

Algorithm 1 Protocole d'annotation multi-experts

Entrées :
— Ensemble d'instances $I = \{i_1, \ldots, i_n\}$
— Ensemble de labels $L = \{l_1, \ldots, l_m\}$
— Ensemble d'experts $X = \{x_1, \ldots, x_p\}$
Sortie : Ensemble d'annotations A
Pour chaque instance $i \in I$ **faire**
  Présenter le prompt correspondant à l'instance
  Pour chaque expert $x \in X$ **faire**
    L'expert x interprète le prompt
    L'expert x sélectionne l'annotation B_i^x telle que $B_i^x \in L$ ou $B_i^x \subseteq L, |B_i| \geq 1$
    L'expert x attribue un degré de confiance :
      $d_i^x = f(\text{compréhension du prompt}) \in [0,1]$
    Enregistrer l'annotation : $a_i^x = (x, B_i^x, d_i^x)$
    Ajouter a_i^x à A
  Fin Pour
Fin Pour
Return A

4.3 Regroupement des annotations

Nous étudions ici différentes façons de regrouper les annotations effectuées par les experts : par *pooling*, par combinaison conjonctive ou par combinaison disjonctive. Rappelons d'abord que chaque annotateur $x \in X$ indique, pour une interaction $i \in I$ donnée, l'ensemble B_i^x des étiquettes qui lui semble pertinent ainsi que son degré de certitude, noté d_i^x.

Pooling. La méthode la plus directe nous semble être le *pooling* des annotations effectuées. Cela correspond à l'interprétation probale telle que décrite dans le modèle à croyance transférables [25] : chaque expert produit un élément de preuve (*piece of evidence*), et la connaissance jointe est simplement l'ensemble des éléments de preuve. La fonction de masse décrivant la connaissance jointe sur l'interaction $i \in I$ est m_i, définie par :

$$m_i(B) = \frac{1}{K_i} \sum_{x \in X, B = B_i^x} d_i^x \quad \text{et} \quad m_i(B) = 0 \text{ sinon.}$$

Les masses sont proportionnelles aux degrés de confiance des experts ; $K_i = \sum_x d_i^x$ est le facteur de normalisation.
C'est la méthode utilisée dans l'exemple 1 (avec des degrés de certitude égaux : $d_i^a = d_a^b > 0$).
Le *pooling* offre ainsi une interprétation claire, tout en gardant la tâche d'annotation simple. Notons qu'à cause de la normalisation, les degrés de certitudes sont considérés de manière relative uniquement. De plus, il n'est pas possible d'itérer le *pooling* (si les annotations d'un autre expert doivent être ajoutées ultérieurement à m_i), car il faut connaître la valeur K_i utilisée afin de re-normaliser. Cette limitation peut être dépassée en conservant cette valeur, ou en utilisant le cadre proposé dans [22] (qui permettrait aussi

d'exprimer le doute sur la fiabilité des experts, ce que nous n'étudions pas ici).
Les éléments focaux de m_i sont les ensembles indiqués par les annotateurs : $\mathcal{S}(m_i) = \{B_i^x \mid x \in X\}$; la complexité est linéaire en le nombre d'annotateurs.

Combinaison conjonctive. La règle de combinaison conjonctive $\otimes_\cap$ [24] et la règle de Dempster $\otimes_\mathfrak{D}$ [8] permettent de combiner des informations exprimées par des fonctions de masse.

$$m^\cap(B) = (m_1 \otimes_\cap m_2)(B) = \sum_{B_1 \cap B_2 = B} m_1(B_1) \times m_2(B_2),$$

$$m^\mathfrak{D}(B) = (m_1 \otimes_\mathfrak{D} m_2)(B) = \begin{cases} 0 & \text{si } B = \emptyset \\ \frac{m_{12}^\cap(B)}{1 - m_{12}^\cap(\emptyset)} & \text{sinon.} \end{cases}$$

La valeur $m^\cap(\emptyset)$ indique le degré de conflit entre m_1 et m_2, et la règle de Dempster ignore ce conflit (les valeurs sont donc renormalisées).
Nous pouvons appliquer ces règles pour combiner les annotations effectuées par les experts sur une même interaction. Nous modélisons alors chaque annotation par une *simple support function* [24] : l'annotation de l'interaction $i \in I$ par l'expert $x \in X$ est la fonction de masse m_i^x telle que :

$$m_i^x(B_i^x) = d_i^x \quad m_i^x(L) = 1 - d_i^x \quad m_i^x(B) = 0 \text{ sinon.}$$

Cela décrit le fait que l'expert x croit à l'ensemble B_i^x avec un degré d_i^x et ne sait pas sinon. Alors, la connaissance jointe concernant l'interaction $i \in I$ est la fonction de masse $m_i = m_i^{x_1} \otimes_\cap m_i^{x_2} \otimes_\cap \cdots \otimes_\cap m_i^{x_k}$.
Notons que si $\forall x, d_i^x = 1$, alors la combinaison conjonctive revient à identifier l'intersection des ensembles d'étiquettes donnés par les experts.
Supposons $d_i^x < 1$ pour tout $x \in X$. Alors, chaque fonction de masses m_i^x a deux éléments focaux : $\mathcal{S}(m_i^x) = \{B_i^x, L\}$. La complexité croit par combinaison : $\mathcal{S}(m_i^x \otimes_\cap m_i^y) = \{B_i^x \cap B_i^y, B_i^x, B_i^y, L\}$. Généralement, $\forall m, \mathcal{S}(m \otimes \cap m_i^x) = \{B \cap B_i^x \mid B \in \mathcal{S}(m)\} \cup \mathcal{S}(m)$, d'où $|\mathcal{S}(m \otimes \cap m_i^x)| \leq 2 \times |\mathcal{S}(m)|$. Par récurrence, nous avons $|\mathcal{S}(m_i^1 \otimes_\cap \cdots \otimes_\cap m_i^k)| \leq 2^k$, la complexité est exponentielle en le nombre d'annotateurs. Cela pourrait être prohibitif mais reste réaliste dans notre contexte où $k = 3$.

Combinaison disjonctive. [24] La combinaison disjonctive considère l'union des ensembles plutôt que leur intersection. La procédure est similaire : les annotations des experts sont exprimées par des *simple support functions* puis combinées par :

$$m^\cup(B) = (m_1 \otimes_\cup m_2)(B) = \sum_{B_1 \cup B_2 = B} m_1(B_1) \times m_2(B_2).$$

Aucun conflit ne peut apparaître avec la règle conjonctive. Notons que si $\forall x, d_i^x = 1$, alors la combinaison disjonctive revient à identifier l'union des ensembles d'étiquettes donnés par les experts.
Supposons $d_i^x < 1$ pour tout $x \in X$. Alors $\mathcal{S}(m_i^x) = \{B_i^x, L\}$. Ainsi, $\mathcal{S}(m_i^x \otimes_\cup m_i^y) = \{B_i^x \cup B_i^y, L\}$. Par récurrence, nous avons $\mathcal{S}(m_i^1 \otimes_\cup \cdots \otimes_\cup m_i^k) = \{\bigcup_{x=1}^k B_i^x, L\}$, d'où $|\mathcal{S}(m_i^1 \otimes_\cup \cdots \otimes_\cup m_i^k)| \leq 2$, la complexité est constante.

Impact de la méthode de regroupement. Les trois méthodes de regroupement étudiées décrivent des processus différents. La méthode « idéale » dépend de la façon dont les experts ont effectué l'annotation.
Le *pooling* est la méthode la plus directe, il ne fait pas d'hypothèse sur le comportement des experts. Elle regroupe simplement leurs annotations en un *pool* d'annotations. Notons que :

- les degrés de confiance éventuellement indiqués par les experts n'y sont considérés que de manière relative,
- les ensembles A tels que $\text{Bel}_i^{\text{Pool}}(A) = 1$ sont les ensembles qui incluent tous les labels identifiés par les experts,
- les ensembles A tels que $\text{Pl}_i^{\text{Pool}}(A) = 1$ sont les ensembles qui contiennent au moins un label identifié par chacun des experts.

En se focalisant sur un sous-ensemble de labels de croyance élevée et de taille minimale, les règles de combinaison peuvent être interprétées comme la première étape d'amélioration des étiquetages : la fusion des annotations revient effectivement à créer une nouvelle annotation jointe.

- La règle conjonctive mène à l'intersection des annotations, elle est adaptée si les experts ont cherché à être exhaustifs (quitte à donner des labels superflus).
- La règle disjonctive mène à l'union des annotations, elle est adaptée si les annotateurs ont cherché à être prudents (quitte à oublier certains labels pertinents).

Nous illustrons l'impact de ces trois méthodes de regroupement des annotations sur deux exemples.

Example 2 (Annotations compatibles) *Considérons l'annotation de l'interaction i avec 3 étiquettes possibles : $L = \{1, 2, 3\}$.*

- *L'expert a dit que $B_i^a = \{1\}$ avec un degré de confiance $d_i^a = 0.6$.*
- *L'expert b dit que $B_i^b = \{1, 2\}$ avec un degré de confiance $d_i^b = 0.9$.*

Le regroupement par *pooling* est direct, associant une masse normalisée de $\frac{d_i^x}{1.5}$ à chaque ensemble B_i^x. Ces masses sont proportionnelles aux degrés de certitude des experts, et mènent aux valeurs listées dans la table 3.
Pour les règles de combinaisons conjonctives et disjonctives, il faut d'abord exprimer les fonctions de masse m_i^x pour chaque expert $x \in X$.

$$m_i^a(\{1\}) = 0.6 \quad m_i^a(L) = 0.4 \quad m_i^a(B) = 0 \text{ sinon,}$$
$$m_i^b(\{1, 2\}) = 0.9 \quad m_i^b(L) = 0.1 \quad m_i^b(B) = 0 \text{ sinon.}$$

Les règles de combinaison disjonctive et conjonctive mènent aux valeurs listées dans la table 3. Puisqu'il n'y a pas de conflit ($m_i^\cap(\emptyset) = 0$), nous avons $m_i^\cap = m_i^\mathfrak{D}$.

Example 3 (Annotations contradictoires) *Considérons l'annotation de l'interaction j avec 3 étiquettes possibles : $L = \{1, 2, 3\}$.*

- *L'expert a dit que $B_i^a = \{1\}$ avec un degré de confiance $d_i^a = 0.6$.*

A	$m_i^{\text{Pool}}(A)$	$\text{Bel}_i^{\text{Pool}}(A)$	$\text{Pl}_i^{\text{Pool}}(A)$	$m_i^{\cap}(A)$	$\text{Bel}_i^{\cap}(A)$	$\text{Pl}_i^{\cap}(A)$	$m_i^{\cup}(A)$	$\text{Bel}_i^{\cup}(A)$	$\text{Pl}_i^{\cup}(A)$
$\emptyset$		0	0		0	0		0	0
$\{1\}$	**0.6**/1.5	0.4	1	0.6	0.6	1		0	1
$\{2\}$		0	0.6		0	0.4		0	1
$\{3\}$		0	0		0	0.04		0	0.46
$\{1,2\}$	**0.9**/1.5	1	1	0.36	0.96	1	0.54	0.54	1
$\{1,3\}$		0.4	1		0.6	1		0	1
$\{2,3\}$		0	0.6		0	1		0	1
$\{1,2,3\}$		1	1	0.04	1	1	0.46	1	1

TABLE 3 – Masses normalisées et degrés de croyance et de plausibilité obtenus par pooling, par fusion disjonctive et par fusion conjonctive pour l'exemple 2.

A	$m_j^{\text{Pool}}(A)$	$\text{Bel}_j^{\text{Pool}}(A)$	$\text{Pl}_j^{\text{Pool}}(A)$	$m_j^{\textcircled{D}}(A)$	$\text{Bel}_j^{\textcircled{D}}(A)$	$\text{Pl}_j^{\textcircled{D}}(A)$	$m_j^{\cup}(A)$	$\text{Bel}_j^{\cup}(A)$	$\text{Pl}_j^{\cup}(A)$
$\emptyset$		0	0		0	0		0	0
$\{1\}$	**0.6**/1.5	0.4	0.4	0.13	0.13	0.22		0	1
$\{2\}$		0	0.6	0.78	0.78	0.87		0	1
$\{3\}$		0	0.6		0	0.09		0	1
$\{1,2\}$		0.4	1		0.91	1		0	1
$\{1,3\}$		0.4	1		0.13	0.22		0	1
$\{2,3\}$	**0.9**/1.5	0.6	0.6		0.78	0.87		0	1
$\{1,2,3\}$		1	1	0.09	1	1	1	1	1

TABLE 4 – Masses normalisées et degrés de croyance et de plausibilité obtenus par pooling, par fusion de Dempster et par fusion conjonctive pour l'exemple 3.

— *L'expert b dit que* $B_i^b = \{2,3\}$ *avec un degré de confiance* $d_i^b = 0.9$.

Le regroupement par pooling est direct, associant une masse de $d_j^x/1.5$ à chaque ensemble B_j^x, ce qui mène aux valeurs présentées dans la table 4.

Pour les règles de combinaison, il faut d'abord exprimer les fonctions de masse m_j^x pour chaque expert $x \in X$.

$$m_j^a(\{1\}) = 0.6 \quad m_j^a(L) = 0.4 \quad m_j^a(B) = 0 \text{ sinon,}$$
$$m_j^b(\{2\}) = 0.9 \quad m_j^b(L) = 0.1 \quad m_j^b(B) = 0 \text{ sinon.}$$

La règle de combinaison disjonctive produit ici un conflit : $m_j^{\cap}(\emptyset) = 0.54 > 0$. La suppression du conflit et renormalisation mènent à la règle de Dempster. Les résultats des règles de Dempster et conjonctives sont listés dans la table 4.

4.4 Qualité des annotations

De nombreuses métriques ont été proposées dans la littérature pour quantifier l'incertitude présente dans une fonction de croyance. Ces métriques visent généralement à mesurer l'imprécision, le conflit ou une combinaison de ces deux dimensions.

Imprécision. Les métriques mesurant l'imprécision se basent sur la cardinalité des éléments focaux (spécificité, entropie de croyance [28, 11, 10, 9]). Dans notre contexte, cela représente l'incertitude structurelle liée aux prompts, qui peuvent être associés à plusieurs étiquettes plausibles.

Nous considérons ici la mesure de spécificité [28] :

$$Sp(m) = \sum_{A \subseteq L} \frac{m(A)}{|A|}.$$

La valeur maximale $Sp(m) = 1$ est atteinte lorsque les annotations de tous les experts sont des singletons (considérant la combinaison disjonctive, il faut en plus que tous les experts soient d'accord sur le même singleton). La valeur minimale $Sp(m) = 1/|\Omega|$ est obtenue lorsque tous les experts affirment conjointement que toutes les étiquettes sont compatibles ($m(L) = 1$).

Conflit. Les métriques mesurant le conflit se basent sur la disjonction des éléments focaux (discorde, dissonance, conflit [16, 29, 15]). Dans notre contexte, elles indiquent donc le désaccord entre les experts.

Nous considérons ici la mesure de dissonance [15] :

$$D(m) = -\sum_{A \subseteq L} m(A) \log_2 \left(1 - \sum_{B \subseteq L} m(B) \frac{|A \cap B|}{|A|}\right)$$

Le terme $\frac{|A \cap B|}{|A|}$ mesure le degré de compatibilité entre deux ensembles de labels, et représente le degré d'accord entre les experts.

Utilisation de ces métriques. Ces deux métriques permettent d'évaluer la qualité des annotations pour chaque interaction. Nous pouvons envisager principalement deux usages :

— **Identifier les annotations problématiques.** Si l'annotation d'une interaction $i \in I$ est décrite

par une fonction de masse fortement dissonante ($D(m_i) > x$, où x est un seuil fixé), cela indique une forte contradiction entre les experts. L'interaction doit alors être réexaminée dans un processus itératif.

— **Évaluer la qualité globale du corpus.** La moyenne des spécificités ou des dissonances permet d'estimer la qualité globale du jeu de données du point de vue de l'incertitude structurelle ou du désaccord entre experts, respectivement.

La capacité à évaluer un jeu de données selon ces deux critères permet non seulement d'analyser l'incertitude présente dans les annotations, mais aussi d'exploiter ces informations pour construire différentes versions du corpus annoté.

5 Sélection de sous-corpus de qualité

Plusieurs stratégies peuvent être envisagées pour produire de nouveaux sous-corpus à partir du jeu de données annoté.

Correction des annotations. Une première approche consiste à corriger les annotations afin de réduire les situations de conflit ou d'imprécision. Toutefois, cette stratégie est en tension avec l'objectif initial de notre protocole, qui vise précisément à préserver l'information portée par l'indétermination annotative. Elle pourrait néanmoins être envisagée dans des cas particuliers, par exemple lorsque la grande majorité des experts s'accorde sur un même ensemble de labels.

Filtrage des interactions. Une seconde stratégie consiste à ne conserver qu'un sous-ensemble des interactions selon des critères de qualité. Par exemple, on peut sélectionner uniquement les instances pour lesquelles la dissonance est inférieure à un seuil donné ou celles dont les annotations présentent un degré élevé de spécificité.

Une telle sélection peut poursuivre deux objectifs : améliorer la cohérence globale du jeu de données utilisé pour l'entraînement ou l'évaluation des modèles, ou isoler des sous-ensembles présentant certaines propriétés particulières, par exemple les interactions dont l'annotation est univoque.

Regroupement des catégories. Une troisième stratégie consiste à regrouper certaines catégories de labels afin de réduire le recouvrement observé dans l'espace catégoriel. Cette opération peut conduire à définir des catégories plus larges pour lesquelles chaque interaction est associée à une étiquette unique.

Ces différentes transformations permettent de générer plusieurs versions du jeu de données, adaptées à différents objectifs d'analyse ou d'apprentissage.

6 Application du cadre proposé au jeu de données « Hackaton »

L'un des intérêts des annotations ensemblistes est qu'elles ne se contentent pas de signaler une indétermination locale : elles permettent aussi de mettre en évidence la structure effective de l'espace catégoriel à partir des combinaisons de labels observées.

Identification des catégories. On observe que les labels 5, 6, 7, 8, 9 et 10 de la table 1 sont souvent regroupés. On trouve par exemple :

— Des interactions pour lesquelles on observe les annotations $\{5, 6\}$, $\{7, 8\}$ ou $\{9, 10\}$ (il s'agit clairement de "*refining*", de "*debugging*" ou de "*reviewing*", mais le code est un mélange de "*own code*" et de "*GPT code*").

— Des interactions pour lesquelles on observe les annotations $\{5, 7, 9\}$ ou $\{6, 8, 10\}$ (il s'agit clairement de "*own code*" ou de "*gpt code*", mais la tache est un mélange de "*refining*", "*debugging*" et "*reviewing*").

— Des interactions pour lesquelles on observe l'annotation $\{5, 6, 7, 8, 9, 10\}$: il s'agit d'une tache mixte pour un code mixte.

La structure de l'espace catégoriel mise en évidence n'est pas une hiérarchie, mais semble s'organiser en treillis (en effet, $\{5\} \subset \{5, 6\} \subset \{5, 6, 7, 8, 9, 10\}$ et $\{5\} \subset \{5, 7, 9\} \subset \{5, 6, 7, 8, 9, 10\}$, mais $\{5, 6\} \not\subseteq \{5, 7, 9\}$ par exemple).

Cette approche permet, en particulier, d'identifier des regroupements d'étiquettes mutuellement exclusifs (tels que la spécificité des annotations de chaque interaction soit égale à 1). Ici, par exemple, les labels $\{5, 6, 7, 8, 9, 10\}$ peuvent être regroupés en une catégorie "code". De cette manière, on construit un jeu de données annoté de manière univoque, en résolvant l'incertitude structurelle identifiée par les experts de manière exacte et sans introduire d'hypothèse arbitraire supplémentaire.

7 Discussion et perspectives

Implications pour l'annotation. Le consensus annotatif ne supprime pas nécessairement l'indétermination structurelle ; il peut au contraire la masquer en imposant une décision exclusive là où plusieurs interprétations demeurent simultanément compatibles. L'annotation mono-label apparaît ainsi moins comme une propriété intrinsèque des données que comme une contrainte méthodologique imposée par le schéma d'annotation.

Perspectives. Plusieurs prolongements de ce travail peuvent être envisagés. D'une part, l'analyse systématique des cooccurrences de labels pourrait permettre de mieux caractériser la structure de l'espace catégoriel et de proposer des schémas d'annotation agrégés. D'autre part, l'intégration du contexte interactionnel constitue une piste importante. Dans ce travail, les prompts ont été annotés de manière isolée afin de mettre en évidence l'indétermination structurelle liée au schéma catégoriel. Cependant, dans certaines situations, l'accès au contexte de l'interaction (prompts précédents, état du code et progression de la tâche) pourrait aider l'annotateur à mieux interpréter la fonction d'un prompt et à préciser l'ensemble des labels plausibles. L'intégration du contexte interactionnel dans le processus d'annotation permettrait de constituer des jeux de données reliant les fonctions des prompts à la progression du travail de l'étudiant et à l'évolution de ses

stratégies d'usage de l'IA.

8 Conclusion

Nous avons montré que, dans le cas étudié, l'annotation fonctionnelle ne peut être modélisée de manière adéquate par une fonction déterministe associant chaque instance à un unique label. L'indétermination observée est structurelle et résulte du recouvrement partiel des catégories du schéma d'annotation. Nous proposons un protocole d'annotation multi-experts permettant une annotation multilabel, dans lequel chaque expert associe à une instance un ensemble de labels plausibles, accompagné d'un degré de confiance reflétant sa certitude quant à l'interprétation du prompt. Pour formaliser cette incertitude épistémique, nous utilisons le formalisme des fonctions de croyance, qui permet de représenter les ensembles de labels proposés par les annotateurs et de quantifier la qualité des annotations en fonction de leur imprécision ou du degré de conflit entre les experts.
Cette approche permet notamment d'identifier des sous-corpus annotés selon différents niveaux de qualité et de mettre en évidence des regroupements pertinents de catégories, révélant la structure effective de l'espace catégoriel. Enfin, en s'inspirant des approches de classification imprécise [18], notre protocole permet de produire des annotations ensemblistes ou crédales, reflétant explicitement l'incertitude associée aux labels des instances. Ces annotations ouvrent la voie à des décisions robustes fondées sur la mesure de l'incertitude, telles que la sélection d'instances ambiguës, la prédiction avec option de rejet ou la génération de prédictions à valeurs multiples.

Références

[1] James Allen and Mark Core. Draft of DAMSL : Dialog act markup in several layers. Technical report, University of Rochester, 1997. Version 2.1.

[2] Lora Aroyo and Chris Welty. Truth is a lie : Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1) :15–24, 2015.

[3] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4) :555–596, 2008.

[4] Petra Saskia Bayerl and Karsten Ingmar Paul. What determines inter-coder agreement in manual annotations ? a meta-analytic investigation. *Computational Linguistics*, 37(4) :699–725, 2011.

[5] Harry Bunt. Multifunctionality in dialogue. *Computer Speech & Language*, 25(2) :222–245, 2011. Language and speech issues in the engineering of companionable dialogue systems.

[6] Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 549–558, Marseille, France, May 2020. European Language Resources Association.

[7] Jean Carletta. Assessing agreement on classification tasks : The kappa statistic. *Computational Linguistics*, 22(2) :249–254, 1996.

[8] Arthur P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2) :325–339, 1967.

[9] Yong Deng. Deng entropy. *Chaos, Solitons & Fractals*, 91 :549–553, 2016.

[10] Didier Dubois and Henri Prade. A note on measures of specificity for fuzzy sets. *International Journal of General System*, 10(4) :279–283, 1985.

[11] Didier Dubois and Henri Prade. The principle of minimum specificity as a basis for evidential reasoning. *Bulletin pour les sous-ensembles flous et leurs applications*, 25 :124–132, 1985.

[12] Didier Dubois and Henri Prade. *Possibility Theory*. Springer US, Boston, MA, 1988.

[13] Hacer Güner and Erkan Er. Ai in the classroom : Exploring students' interaction with chatgpt in programming learning. *Education and Information Technologies*, pages 1–27, 2025.

[14] Nancy Ide and Jean Véronis, editors. *Corpus Annotation : Linguistic Information from Computer Text Corpora*. Kluwer Academic Publishers, Dordrecht, 1998.

[15] George J Klir and Behzad Parviz. A note on the measure of discord. In *Uncertainty in Artificial Intelligence*, pages 138–141. Elsevier, 1992.

[16] George J Klir and Arthur Ramer. Uncertainty in the dempster-shafer theory : a critical re-examination. *International Journal of General System*, 18(2) :155–166, 1990.

[17] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

[18] Vu-Linh Nguyen, Haifei Zhang, and Sébastien Destercke. Credal ensembling in multi-class classification. *Machine Learning*, 114(1) :19, 2025.

[19] Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7 :677–694, 2019.

[20] Barbara Plank. The "problem" of human label variation : On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[21] Massimo Poesio and Ron Artstein. The reliability of anaphoric annotation, reconsidered : taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II : Pie in the Sky*, CorpusAnno '05, page 76–83, USA, 2005. Association for Computational Linguistics.

[22] Pierre Pomeret-Coquot. Modeling and updating uncertain evidence within belief function theory. *International Journal of Approximate Reasoning*, 182 :109428, 2025.

[23] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Sebastopol, CA, 2012.

[24] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[25] Philippe Smets and Robert Kennes. The transferable belief model. *Artificial intelligence*, 66(2) :191–234, 1994.

[26] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification : An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3) :1–13, July 2007.

[27] Alexandra Uma, Barbara Plank, Dirk Hovy, Tommaso Fornaciari, and Massimo Poesio. Learning from disagreement : A survey. *Journal of Artificial Intelligence Research*, 72 :1385–1470, 2021.

[28] Ronald R Yager. Entropy and specificity in a mathematical theory of evidence. *International Journal of General System*, 9(4) :249–260, 1983.

[29] Ronald R Yager. On the Dempster-Shafer framework and new combination rules. *Information sciences*, 41(2) :93–137, 1987.