Identification automatique des oronymes dans un corpus de récits d'exploration des Alpes

Xiao Ma

Identification automatique des oronymes dans un corpus de récits d'exploration des Alpes

Xiao
MA

Sous la direction de Olivier KRAIF et Samia OUNOUGHI

Laboratoire : LIDILEM

UFR LLASIC
Département I3L

Mémoire de master 2 mention Sciences du langage - 20 crédits

Parcours : Industries de la langue, orientation professionnelle

Année universitaire 2022-2023

Identification automatique des oronymes dans un corpus de récits d'exploration des Alpes

Xiao
MA

Sous la direction de Olivier KRAIF et Samia OUNOUGHI

Laboratoire : LIDILEM

UFR LLASIC
Département I3L

Mémoire de master 2 mention Sciences du langage - 20 crédits

Parcours : Industries de la langue, orientation professionnelle

Année universitaire 2022-2023

Remerciements

Je tiens à exprimer ma plus sincère gratitude envers de nombreuses personnes pour leur contribution précieuse à la réussite de ce stage et à la rédaction de ce mémoire.

Tout d'abord, j'ai l'honneur de remercier profondément mes deux encadrants, M. Olivier Kraif et Mme Samia Ounoughi pour leur encadrement. Leur expertise et leurs conseils éclairés ont été fondamentaux tout au long de mon stage. Leur soutien et leurs orientations ont contribué de manière significative à la réalisation de ce travail.

Je souhaite également exprimer ma reconnaissance à mes amis. Florine, tes conseils et ta présence m'ont accompagné tout au long de mon parcours de master, et je suis reconnaissant pour chaque moment partagé. 陳禹, tes connaissances et tes discussions passionnantes sur la linguistique ont enrichi mon apprentissage de manière inestimable.

Enfin, je souhaite adresser un merci tout spécial à ma famille et à mes meilleurs amis, 王沁, 鍾銘輝, 張婷. Votre amour et vos encouragements sont mes plus grands trésors, merci d'avoir toujours été à mes côtés.

DÉCLARATION ANTI-PLAGIAT

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

PRENOM : Xiao……

NOM : MA……

DATE : 20/07/2023……

SOMMAIRE

Introduction

Dans ce mémoire, nous nous intéressons à l'entraînement et l'évaluation d'un modèle d'apprentissage profond en nous appuyant sur BERT (Devlin et al., 2019), un modèle pré-entraîné basé sur l'architecture Transformer (Vaswani et al., 2017). Ce modèle est spécialisé sur la détection d'un type de nom propre particulier, les oronymes[1], dans un corpus en anglais qui se compose de récits d'exploration des Alpes. Les données du corpus proviennent de la revue géographique *The Alpine Journal*, et contiennent au total 1,4 millions de mots.

Dans des travaux précédents, Samia Onoughi, ma tutrice de stage, avait tenté d'utiliser le langage de requête du Lexicoscope (Kraif & Diwersy, 2012)[2] pour identifier des occurrences d'oronymes dans ce corpus. Étant donné que la diversité morphosyntaxique des oronymes augmente considérablement le nombre de requêtes nécessaires pour obtenir un inventaire fiable, celle-ci entraîne souvent de nombreux silences. En outre, les structures morphosyntaxiques que l'oronyme peut revêtir correspondent parfois à des syntagmes qui ne sont pas des oronymes, générant ainsi beaucoup de bruit dans les résultats. Il est donc essentiel de mettre au point un outil capable de repérer efficacement les oronymes tout en minimisant les silences et les bruits en fonction des besoins de notre tâche, ce qui constitue notre problématique.

Dans cette perspective, notre démarche s'est orientée vers l'exploration d'une approche d'optimisation (*fine-tuning*) basée sur des modèles dérivés du Transformer, en vue de résoudre notre problématique. Nous nous sommes interrogés sur l'efficacité potentielle de l'application des modèles de traitement du langage naturel (TAL) dans la détection d'oronymes en contexte. Cette décision a émergé à la suite de l'analyse des questions initiales de notre recherche, à savoir le choix des modèles et le traitement des données. De plus, l'hypothèse sous-jacente à notre démarche, à savoir l'utilisation d'une approche basée sur les modèles Transformers pour la détection des entités nommées, a son importance.

Notre point de départ réside dans la liste d'oronymes hors contexte constituée et validée manuellement par Samia Ounoughi. Nous avons l'hypothèse que l'utilisation d'une approche basée sur les modèles Transformers pour la détection des entités nommées pourrait être adaptée à notre cas. L'idée est d'entraîner un classifieur sur ces données en exploitant les avantages des modèles pré-entraînés. Cependant, la question de l'ajustement du modèle pour qu'il soit efficace

[1] Oronyme : (Cartographie, Géographie, Toponymie) Nom porté par un élément de relief : une montagne, une chaîne de montagnes, une colline, une montagne sous-marine, etc. (Comité français de cartographie, 1990)
[2] Source : http://phraseotext.univ-grenoble-alpes.fr/Lexicoscope_2.0/

dans la tâche spécifique de détection d'oronymes en contexte est devenue le cœur de notre démarche.

Dans un premier temps, nous reviendrons sur le contexte dans lequel s’est déroulé ce stage ainsi que sur la mission dans le chapitre 1. Ensuite, nous développerons les différentes méthodologies abordées dans le chapitre 2. Nous procéderons ensuite à l'analyse des résultats obtenus, exposée dans le chapitre 3. Enfin, nous ouvrirons la voie aux perspectives envisageables tout en offrant une réflexion personnelle et professionnelle dans le dernier chapitre.

Chapitre 1. Contexte et mission

Dans ce premier chapitre, entièrement dédié à la contextualisation de mon stage, nous commencerons par présenter les informations essentielles concernant mon stage, en particulier sur la façon dont il s'est déroulé. Nous ferons ensuite une brève présentation de l'organisme d'accueil.

Par la suite, nous explorerons en détails les différentes ressources que nous avons utilisées tout au long du stage. Cela inclura une description complète du corpus de données que nous avons manipulé, en mettant en avant ses caractéristiques, sa provenance, ainsi que les méthodes de collecte et d'annotation utilisées. De plus, nous fournirons une présentation détaillée des autres ressources pertinentes qui ont été utilisées dans le cadre de mes travaux.

Finalement, nous aborderons la problématique spécifique que l'on a soulevée pendant le stage et également l'objectif global qui a été fixé pour répondre à cette problématique.

1. Contexte du stage

Cette mission s'inscrit dans le cadre de mon stage de fin d'études au sein du programme de Master en Sciences du Langage, avec une spécialisation dans les Industries de la Langue. Mon stage est co-encadré par Mme Samia Ounoughi, maîtresse de conférences à l'Université Grenoble Alpes[3], et M. Olivier Kraif, professeur à l'Université Grenoble Alpes. Ce stage fait partie du projet NOMINALP[4], financé par CerCog@UGA[5], et accueilli par le Laboratoire de Linguistique et Didactique des Langues Étrangères et Maternelles (ci-après Lidilem[6]).

1.1. Organisme d'accueil

Le Lidilem est fondé en 1987, rattaché à l'Université Grenoble Alpes. Il compte à l'heure actuelle une soixantaine de chercheurs permanents et environ 70 doctorants. Les travaux de recherche menés au sein du Lidilem couvrent divers domaines des sciences du langage, tels que les descriptions linguistiques, la sociolinguistique, l'acquisition des langues, la constitution et l'exploitation de corpus, la didactique des langues, le traitement automatique des langues (TAL), ainsi que l'étude des nouvelles formes d'interaction suscitées par les usages numériques.

[3] Source : https://www.univ-grenoble-alpes.fr/
[4] NOMINALP : Name of mountains in the Alps (Noms des montagnes dans les Alpes)
[5] Source : https://cercog.univ-grenoble-alpes.fr/
[6] Source : https://lidilem.univ-grenoble-alpes.fr/

1.2. Problématique

Le nom propre constitue une sous-catégorie de noms qui a fait l'objet d'études approfondies de la part des logiciens et des linguistes. De nos jours, un consensus émerge au sein de la littérature académique en linguistique, selon lequel la grammaire du nom propre (Gary-Prieur, 1994) ne fait pas de distinction entre les différents types selon leur structure morphosyntaxique ou la nature de leur référent, et il est fréquent que les résultats obtenus soient lacunaires voire incorrects (Ounoughi, 2024). Cependant, une contribution significative peut être trouvée dans le domaine de l'onomastique, une sphère particulière qui fournit des informations extrêmement précises sur les noms propres individuels, en ce qui concerne leur origine, leur évolution, et plus encore. Toutefois, en raison de l'hétérogénéité des noms propres, des investigations supplémentaires sont requises pour mieux appréhender le fonctionnement des noms propres en tant que catégorie.

Actuellement, ce projet se concentre sur une catégorie spécifique de noms propres, à savoir les oronymes, qui font référence à des formations géologiques, principalement des montagnes. Dans cette perspective, l'objectif est d'explorer les vastes ensembles de données afin d'identifier et d'analyser les occurrences des oronymes dans le discours. À cette fin, un corpus de récits d'exploration des Alpes en anglais datant du 19ème siècle a été constitué par Samia Ounoughi (Ounoughi, 2023) et a été ensuite examiné au moyen de l'outil LEXICOSCOPE (Kraif & Diwersy, 2012), qui est un outil d'exploration de corpus spécialisé à l'exploration de profils combinatoires de mots ou d'expressions, en se basant sur les dépendances syntaxiques.

Les travaux de recherches de Samia Ounoughi sur le nom propre montrent que l'apparition d'un oronyme implique un ensemble complexe de processus cognitifs et linguistiques, comprenant l'exploration, l'identification, la découpe, les tentatives de désignation et la perpétuation du nom. À partir d'un vaste corpus de textes renfermant de nombreux oronymes, l'analyse linguistique de corpus permet d'observer ces phénomènes cognitifs et linguistiques dans leur contexte. L'objectif est de mettre en lumière les caractéristiques communes ainsi que les spécificités de ces oronymes au sein de leurs contextes, et d'expliciter les mécanismes qui influent sur notre perception de l'espace, de sa dénomination, de ses motifs de dénomination et des méthodes utilisées pour le nommer. Ces approches visent à dévoiler comment la relation entre le langage et l'espace permet à l'être humain de donner une signification à l'espace afin de mieux l'explorer, l'habiter, voire le transformer.

À présent, l'un des freins majeurs du projet de recherche réside dans la méthode peu efficace utilisée pour recenser les oronymes au moyen de requêtes sur la plateforme Lexicoscope. Cette situation découle de diverses raisons. Tout d'abord, la morphosyntaxe des oronymes présente une grande diversité de variations, ce qui engendre une augmentation significative du nombre de requêtes ainsi qu'une abondance de réponses négatives. De plus, la morphosyntaxe des oronymes est souvent similaire à celle de syntagmes nominaux qui ne correspondent pas à des oronymes, entraînant ainsi la présence d'un volume substantiel de résultats non pertinents. Ma tutrice, Mme Samia Ounoughi, étant linguiste, souhaite avoir une méthode efficace pour extraire les occurrences des oronymes dans leur contexte à partir du corpus. Cependant, les requêtes morphosyntaxiques seules ne se révèlent pas suffisantes pour réaliser cette tâche. En tant que stagiaire en ingénierie du TAL, ma responsabilité est de trouver une solution qui permettra à ma tutrice d'accéder facilement à ces occurrences en contexte. Pour y parvenir, nous avons décidé de procéder à une annotation préalable du corpus.

1.3. Objectif

L'objectif de ce stage est de repérer les oronymes au sein du corpus en développant un outil de détection d'entités nommées basé sur l'apprentissage automatique. Cette démarche aboutira à l'élaboration d'une liste fiable d'oronymes dans le cadre du projet de recherche ainsi qu'à l'élaboration d'un corpus annoté permettant d'explorer les oronymes en contexte. Une liste d'oronymes hors contexte, préalablement extraite manuellement du corpus, est déjà disponible.

La première étape consiste à associer ces oronymes à l'ensemble du corpus afin de les situer dans leurs contextes et de créer un ensemble de données d'entraînement basé sur cette démarche. Par la suite, nous procéderons à l'entraînement d'un modèle supervisé visant à détecter les oronymes. Un modèle supervisé est un type de modèle d'apprentissage automatique qui nécessite un ensemble de données préalablement étiquetées pour son entraînement.

Dans le cadre de l'apprentissage supervisé, le modèle ajuste ses paramètres en comparant ses prédictions aux étiquettes réelles des données d'entraînement. Cette comparaison permet au modèle de s'ajuster progressivement afin d'améliorer ses performances. Une fois que le modèle montre de bonnes performances sur les données d'entraînement, il peut être utilisé pour prédire de nouvelles données non vues.

Par la suite, le corpus nouvellement obtenu à partir de l'étape précédente sera subdivisé en ensembles distincts, à savoir l'ensemble d'entraînement, l'ensemble de validation et

l'ensemble de test, qui serviront à la formation du modèle. Ce processus est expliqué en détails dans la section consacrée à la division des données dans le chapitre II.

La deuxième phase implique l'optimisation de modèles déjà pré-entraînés sur nos données d'entraînement. Par la suite, divers modèles génératifs ainsi que des modèles pré-entraînés pour la détection des toponymes[7] seront évalués. Une évaluation comparative sera réalisée en testant différents modèles et configurations de paramètres, afin de sélectionner la méthode la plus efficace. Le modèle retenu sera ensuite utilisé pour annoter l'intégralité du corpus, pour importer *in fine* cette nouvelle version annotée dans la plateforme Lexicoscope.

2. Base de travail

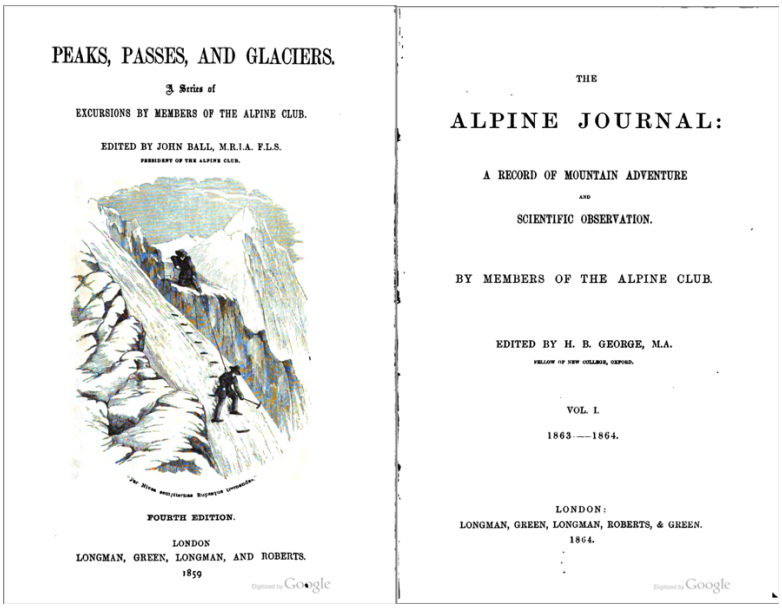
Dans cette deuxième section, nous présenterons en détails les ressources qui ont été mises à ma disposition au début de mon stage. Ces ressources sont essentielles pour le démarrage et la réalisation de mes travaux. Elles comprennent tout d'abord une description de la revue d'alpinisme, ainsi qu'un corpus constitué à partir des textes de cette revue. Outre le corpus, une liste de 1221 oronymes a également été mise à ma disposition. En plus de ces ressources linguistiques, nous décrirons l'architecture du Transformer, ainsi que celle des deux modèles de traitement du langage naturel qui ont été pré-entraînés sur Transformer, à savoir BERT et DistilBERT (Sanh et al., 2020). Enfin, nous présenterons brièvement le tutoriel officiel de DistilBERT.

Ces ressources constituent la base de mon travail. Je décrirai leur utilité et leur pertinence pour atteindre les objectifs du stage dans les sections suivantes.

2.1. The Alpine Journal

The Alpine Journal (ci-après *AJ*) a été fondé au 2 mars 1863, mais a été initialement publié sous le nom de *Peaks, Passes, and Glaciers* en 1858 par le premier club d'alpinisme au monde, l'*Alpine Club of London*[8] (Club Alpin de Londres), qui a été créé en 1857. Cette revue est reconnue comme étant la plus ancienne revue d'alpinisme au monde (Stephen, 2004).

[7] Toponyme : est un nom de lieu ; Source : https://dictionnaire.lerobert.com/guide/toponymes
[8] Source : http://www.alpine-club.org.uk/



PEAKS, PASSES, AND GLACIERS.

A Series of

EXCURSIONS BY MEMBERS OF THE ALPINE CLUB.

EDITED BY JOHN BALL, M.R.I.A. F.L.S.

FOURTH EDITION.

LONDON
LONGMAN, GREEN, LONGMAN, AND ROBERTS.
1859

THE

ALPINE JOURNAL:

A RECORD OF MOUNTAIN ADVENTURE
AND
SCIENTIFIC OBSERVATION.

BY MEMBERS OF THE ALPINE CLUB.

EDITED BY H. B. GEORGE, M.A.

VOL. I.

1863—1864.

LONDON:
LONGMAN, GREEN, LONGMAN, ROBERTS, & GREEN.
1864.

Figure 1. Peaks, Passes and Glaciers (à gauche) et The Alpine Journal (à droite)

L'*AJ* a été créé dans le but de couvrir un large éventail des expéditions d'alpinisme de toutes sortes à travers le monde, et pas seulement dans les Alpes. Plusieurs sujets ont été traités dans cette revue, notamment des histoires de l'expédition, des aventures, de l'art, de la littérature, de la géographie, de l'histoire, de la géologie, de la médecine, de l'éthique, de l'environnement montagnard, etc. Toutes ces nouvelles connaissances scientifiques et géographiques, issues de différentes sources, ainsi que les nouveaux livres sur les Alpes, ont revêtu une importance significative pour la communauté des alpinistes de l'époque. En effet, les sociétés d'alpinisme étaient encore très minoritaires à ce moment-là, et le *London Alpine Club* était le seul de ce genre à exister jusqu'en 1862, avant la création des sociétés d'alpinisme autrichiennes, suisses et italiennes.

Aujourd'hui, la grande majorité des éditions papier précédemment publiées ont été numérisées et sont maintenant accessibles gratuitement sur le site officiel de l'*AJ*. De plus, de nombreuses universités et bibliothèques du monde entier ont également numérisé leurs propres collections d'éditions papier de l'*AJ*, les rendant disponibles en accès libre et en téléchargement gratuit en ligne. La version numérisée que nous avons obtenue pour le projet provient de la Bibliothèque nationale autrichienne (*Österreichische Nationalbibliothek*[9]), numérisée par l'équipe de Google, et comprend un total de 12 109 pages, avec une taille de fichier de 596,6

[9] Source : https://www.onb.ac.at/

Mo au format PDF. Dans la prochaine section, "Le corpus", nous fournirons une introduction plus détaillée du corpus.

2.2. Le corpus

Le corpus que nous utilisons dans cette tâche a été constitué précédemment par l'un de mes tuteurs, Mme Samia Ounoughi en collaboration avec ses anciens étudiants, Anitaa Paramanantham, Célia Marion et Judith Chambre. Ce travail fait partie du projet DÉMARRE SHS ! est le wp3 de DATA @ UGA[10], officiellement lancé en 2017. Ce projet a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-15-IDEX-02. Le projet vise initialement à construire une communauté de recherche sur les données des sciences humaines et sociales (SHS) à Grenoble, puis à répondre aux besoins urgents en matière d'hébergement, d'exposition, d'exploitation et de pérennisation des données SHS.

Le corpus se compose de 384 fichiers au format XML. Ce format a été conçu initialement pour résoudre les enjeux de l'édition électronique à grande échelle et qui joue désormais un rôle crucial dans de nombreuses interactions de données sur le Web et au-delà. Ces fichiers sont structurés dans 21 dossiers organisés par année, comme indiqué dans le tableau 1. De plus, l'histogramme présenté sur la figure 2 offre une visualisation de la répartition des fichiers dans le temps.

Année[ab]	58	60	64	66	67	70	72	74	76	78	80	82	84	86	88	89	91	93	95	97	99
Nombre	16	1	36	15	8	14	27	29	32	22	31	21	17	10	19	11	18	12	11	11	17

Tableau 1. Nombre de fichiers par année

[10] Source : https://demarreshs.hypotheses.org/

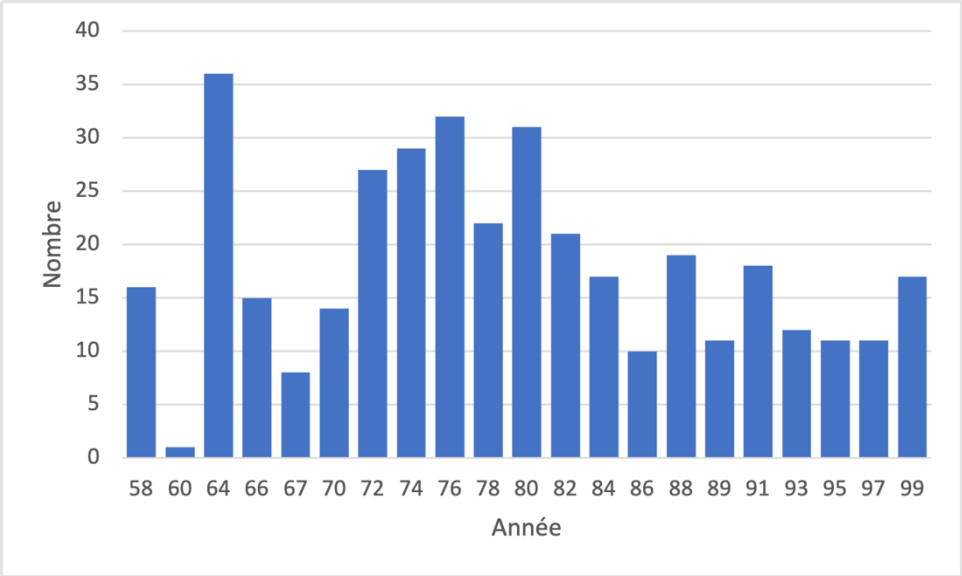


Figure 2. Répartition des fichiers par année de publication

a. Toutes les années débutent par "18", mais "18" a été omis en raison de contraintes d'espace.

b. Seules les années 1858, 1860, 1864, 1866, 1867, 1870 et 1872 contiennent des fichiers qui ont été vérifiés manuellement.

Les données du corpus proviennent d'articles publiés dans l'*AJ*, au cours des années mentionnées dans le tableau 1. Étant donné que les créateurs du corpus possédaient déjà une version numérisée de l'*AJ* au format PDF, et afin d'extraire tous les textes de manière efficace, il a été décidé d'utiliser la technologie de la reconnaissance optique de caractères (OCR[11]). Il a ensuite été révisé et a été annoté manuellement en respectant la version 1.0 de la norme TEI[12], une initiative normalisée de codage de texte lancée conjointement en 1987 par l'Association for Computers and the Humanities[13], l'Association for Computational Linguistic[14] et l'Association for Literary and Linguistic Computing[15].

La TEI a pour objectif d'annoter et d'encoder des textes en vue de faciliter leur échange, leur partage et leur traitement dans l'environnement numérique. Elle consiste en un ensemble de spécifications de balisage qui définit comment décrire et baliser différents types d'informations dans un texte, tels que la structure du texte, la sémantique, la langue, les annotations, les références, etc. Grâce à l'utilisation du balisage TEI, il est possible de convertir les textes dans un format lisible par l'ordinateur, permettant ainsi leur analyse, leur recherche, leur traitement et leur présentation dans un environnement numérique. Cela facilite également

[11] OCR : Optical character recognition
[12] TEI : Text Encoding Initiative; Source : https://tei-c.org/ns/1.0
[13] Source : https://ach.org/
[14] Source : https://www.aclweb.org/portal/
[15] ALLC : Actuellement appelée European Association for Digital Humanities; Source : https://eadh.org/

la comparaison, la mise en relation et le partage des textes avec d'autres sources, contribuant ainsi au développement et à la promotion de la recherche en sciences humaines numériques.

Parmi tous ces 384 fichiers XML annotés, chaque fichier individuel représente un chapitre d'un article figurant dans le fichier PDF d'origine. Pour chaque fichier XML, il existe deux parties principales, la première étant les métadonnées et la seconde le contenu.

Dans un premier temps, la structure de la section des métadonnées comprend les informations de base suivantes : le titre, le bailleur de fonds, le responsable principal du projet, l'annotateur, le vérificateur, le concepteur du schéma TEI, ainsi que leurs affiliations respectives et l'année de l'annotation. Une deuxième partie importante porte sur les métadonnées de la revue elle-même, telles que le titre de la revue, le sous-titre, le titre de l'article, l'auteur de l'article, la maison d'édition, le lieu de publication, l'année de publication, l'éditeur, le numéro de volume et la pagination. Toutes les informations ci-dessus constituent un ensemble complet de métadonnées. Il existe également une autre partie qui indique le sujet, l'année et la langue de l'article du chapitre. Toutes ces informations constituent un ensemble de métadonnées.

Quant à la structure du contenu, elle est entièrement basée sur la mise en forme de chaque page du fichier PDF d'origine. Dans l'ensemble, chaque page de l'édition papier est annotée en fonction de son contenu. Par exemple, chaque paragraphe est encadré par une paire de balises `<p>` et `</p>` pour indiquer un paragraphe. De même, pour chaque phrase, si un saut de ligne apparaît dans le livre, une balise `<lb/>` est utilisée pour signaler un saut de ligne dans la version annotée. De plus, les titres, les numéros de page et les notes ont également été annotés.

2.3. Liste des oronymes

Après la création réussie du corpus, une liste[16] de 1221 oronymes valides a été obtenue en utilisant différentes méthodes de requête sur la plateforme Lexicoscope. Ces oronymes sont divisés en deux catégories principales : ceux qui contiennent le déterminant anglais « *the* » devant chaque oronyme, avec un total de 1148 dans cette catégorie, et ceux qui ne le contiennent pas, avec un total de 73.

Parmi ces oronymes, on retrouve principalement des noms propres liés aux termes topographiques tels que *Glacier, Col, Val, Vallée, Valley, Mont, Monte, Pic, Bec, Pass, Passo* en français, italien et anglais. Il y a aussi des noms propres en allemand liés aux termes topographiques qui se terminent par des suffixes en allemand tels que *-horn, -thal, -berg, -joch,*

[16] Voir Annexe n°1 à la page 82

-hörner, -stein, -spitz. Une autre catégorie comprend les noms propres simples comme *the Alps*, *the Sardona*, *the Great St. Bernard*, etc.

2.4. Introduction sur les modèles Transformer, BERT et DistilBERT

2.4.1. Les Transformers

Le Transformer (Vaswani et al., 2017) est une architecture de réseau neuronal conçue par une équipe de chercheurs de Google et de l'Université de Toronto en 2017. Il a été initialement conçu pour résoudre le problème de la précision limitée des modèles Seq2seq (Sutskever et al., 2014) lorsqu'ils sont appliqués à des séquences longues dans la traduction automatique. L'émergence du Transformer a constitué une solution très satisfaisante à ce problème, voire une étape révolutionnaire dans le développement du domaine du traitement du langage naturel (TAL).

Le processus d'entraînement du Transformer utilise deux ensembles de données : le standard WMT 2014[17] English-German et le plus grand WMT 2014 English-French. Le premier contient 4,5 millions de paires de phrases et le second 36 millions de phrases. Huit GPUs[18] de type NVIDIA P100 ont été employés pour l'entraînement. Il existe deux versions du Transformer : le modèle de base et le modèle grand. La version de base subit un processus d'entraînement sur une durée de 100 000 étapes, équivalent à environ 12 heures, avec chaque étape s'exécutant en environ 0,4 seconde. D'autre part, la version grande est entraînée sur 300 000 étapes, nécessitant environ 3,5 jours au total, avec chaque étape s'étalant sur environ 1,0 seconde.

[17] Source : https://www.statmt.org/wmt14/translation-task.html
[18] GPU : unité de traitement graphique

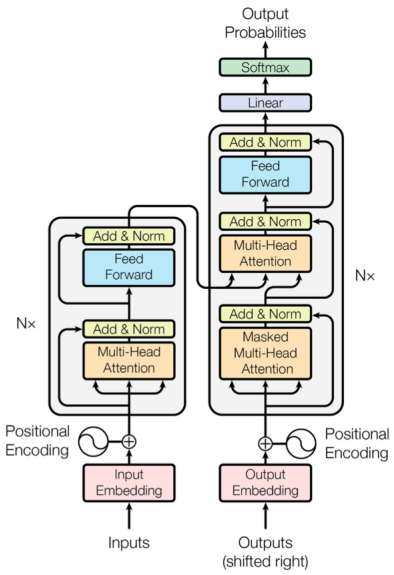


Figure 3. L'architecture du Transformer (Vaswani et al., 2017), cette illustration est présentée dans l'article original

La caractéristique essentielle du Transformer est le mécanisme d'auto-attention (*self-attention*), ce qui permet au modèle de prendre en compte les informations provenant d'une position lorsqu'il calcule d'autres positions, ce nouveau mécanisme d'attention lui permettant donc de prendre en compte les relations dans des séquences plus longues et de traiter des structures linguistiques plus complexes.

Le modèle Transformer se compose de deux composants principaux, comme le montre la figure 3. À gauche se trouve l'encodeur, qui est structuré en six couches identiques, chacune comprenant deux sous-couches distinctes. La première sous-couche intègre un mécanisme d'auto-attention à têtes multiples, tandis que la seconde est un réseau *feed-forward* entièrement connecté.

De l'autre côté, nous avons le décodeur, qui partage également une structure en six couches. Cependant, il incorpore une sous-couche supplémentaire qui permet l'application d'une attention à têtes multiples sur la sortie obtenue de l'ensemble de l'encodeur.

Grâce à cette spécificité, le Transformer peut se concentrer simultanément sur des éléments différents de la séquence d'entrée, afin de capturer efficacement les informations sémantiques globales au cours du processus d'encodage et de décodage. Les modèles précédents, par exemple, les réseaux neuronaux récurrents (RNN) et les réseaux neuronaux convolutifs (CNN) souffrent au contraire de problèmes de dépendance à long terme lorsqu'ils traitent de longues chaînes de données, ce qui entraîne un manque d'exhaustivité dans le transfert

d'informations. L'émergence des Transformers a révolutionné l'état de l'art en permettant une attention multi-tête qui permet de capturer des informations à longue portée dans les séquences. Cette avancée ouvert la voie à des modèles encore plus puissants, tels que BERT, qui se fondent sur ces bases pour atteindre une compréhension encore plus profonde et précise du langage.

2.4.2. BERT

BERT (Devlin et al., 2019), acronyme de Bidirectional Encoder Representations from Transformers, est un modèle de traitement du langage naturel (TAL) pré-entraîné, développé par l'équipe de Google en 2018. Tel que son nom l'indique, BERT est un modèle bidirectionnel qui repose sur l'architecture Transformer, plus précisément il est composé uniquement d'un encodeur. Par conséquent, il hérite du mécanisme d'auto-attention caractéristique de Transformer, reconnu pour sa capacité à saisir les informations contextuelles dans le texte.

Contrairement aux modèles de langage conventionnels, qui sont généralement unidirectionnels, BERT analyse les contextes à la fois à gauche et à droite de chaque mot dans une phrase. Cette approche bidirectionnelle permet à BERT de saisir plus efficacement la sémantique et le contexte des mots, ce qui se traduit par une meilleure représentation et compréhension du langage naturel.

Initialement, BERT comportait deux variantes basées sur la langue anglaise : Base et Large. Le modèle Base est constitué de 12 encodeurs, chacun avec 12 têtes d'auto-attention bidirectionnelle, ainsi que 110 millions de paramètres. Le modèle Large, quant à lui, présente 24 encodeurs, chacun équipé de 16 têtes d'auto-attention bidirectionnelle, et compte 340 millions de paramètres. Pour leur entraînement, les deux modèles ont utilisé des données issues de deux corpus distincts : BookCorpus (800 millions de mots) (Zhu et al., 2015) et English Wikipedia (2,5 milliards de mots).

Le cadre de BERT est constitué de deux phases distinctes : la pré-entraînement et le réglage fin (*fine-tuning*). La phase de pré-entraînement vise à obtenir une représentation linguistique générique grâce à un apprentissage non supervisé à grande échelle. Cette phase comprend deux tâches principales : la Modélisation de Langage Masqué (MLM) et la Prédiction de la Phrase Suivante (NSP[19]). Dans la tâche de Modélisation de Langage Masqué, le modèle est chargé de prédire les 15% de mots qui ont été masqués aléatoirement dans le texte. En parallèle, la tâche de Prédiction de la Phrase Suivante consiste à déterminer si deux phrases données sont consécutives dans le texte original, deux marqueurs spéciaux [CLS] et [SEP] ont

[19] NSP : Next Sentence Prediction

été employés pour marquer le début et la fin d'une phrase. Ces marqueurs permettent à BERT d'apprendre à discerner la continuité entre les phrases et de mieux comprendre la structure du texte.

BERT peut être utilisé dans diverses tâches de TAL après avoir été entraîné sur des grands corpus. Les tâches de réglage fin peuvent inclure la classification de paires de phrases, la classification de phrases simples, les systèmes de question-réponse, la reconnaissance d'entités nommées (NER [20]), et d'autres encore. BERT a atteint des performances exceptionnelles sur 11 évaluations distinctes en TAL, se démarquant particulièrement dans le cadre du référentiel GLUE (Wang et al., 2019)[21] avec un score remarquable de 80,5 %, représentant une amélioration absolue significative de 7,7 %. De même, dans la tâche de précision MultiNLI (Williams et al., 2018)[22], BERT a enregistré un score de 86,7 %, présentant une amélioration absolue notable de 4,6 %. Ces résultats marquent une étape majeure dans l'évolution du traitement du langage naturel, démontrant l'efficacité et la pertinence de l'approche bidirectionnelle et du mécanisme d'auto-attention de BERT.

En outre, dans le domaine de l'apprentissage automatique, la disponibilité de données d'entraînement suffisantes constitue souvent un obstacle à surmonter. C'est ce qui distingue ces modèles : ils sont pré-entraînés sur un grand volume de textes, ce qui leur permet d'acquérir une compréhension générale de la langue. Un avantage important de ce pré-entraînement est que nous pouvons affiner ces modèles en utilisant des ensembles de données plus petits. De cette manière, ces modèles peuvent être adaptés à des tâches spécifiques même si la quantité de données d'entraînement spécifiques est limitée.

2.4.3. DistilBERT

Étant donné que les variantes du modèle Transformer continuent de produire des résultats de pointe dans diverses tâches de traitement du langage naturel (TAL), le nombre croissant de paramètres ainsi que les besoins en matériel de calcul plus puissant augmentent de manière presque exponentielle. En conséquence, le processus d'entraînement devient de plus en plus coûteux pour les chercheurs. De ce fait, l'équipe de Hugging Face[23] a introduit DistilBERT (Sanh et al., 2020) en 2019.

[20] NER : Named Entity Recognition
[21] GLUE : General Language Understanding Evaluation (Évaluation de la compréhension générale de la langue)
[22] MultiNLI : Multi-Genre Natural Language Inference (Inférence en langage naturel multi-genre)
[23] Hugging Face est une organisation de recherche en intelligence artificielle et une communauté open source réputée, basée à New York et fondée en 2016, qui se concentre sur le développement et le partage de modèles et d'outils de TAL; Source : https://huggingface.co/

DistilBERT est une version allégée de BERT Base, conçue pour préserver des performances élevées tout en offrant plusieurs avantages supplémentaires. En plus de conserver des performances comparables, DistilBERT présente des améliorations substantielles dans d'autres domaines clés. Par exemple, il se caractérise par des temps d'entraînement plus courts, ce qui permet d'accélérer le processus de développement de modèles. De plus, en raison de sa taille réduite, il nécessite moins de ressources computationnelles, ce qui en fait une option plus économique pour de nombreux projets de TAL. Par conséquent, l'utilisation de DistilBERT offre non seulement une solution efficace en termes de performances, mais aussi des avantages pratiques en termes de temps et de coûts d'entraînement.

La compression du modèle (Bucila et al., 2006) et la distillation de connaissances (Hinton et al., 2015) fait référence au processus de compresser un modèle élève pour reproduire le fonctionnement d'un modèle enseignant qui est plus grand. Il implique l'extraction des connaissances d'un modèle vaste et complexe, suivi du transfert de ces connaissances vers un modèle plus compact et léger. Cela vise à réduire la taille du modèle ainsi que la charge en ressources informatiques, tout en préservant des performances de haut niveau. Cette approche trouve des similitudes avec les méthodes d'apprentissage et de transfert de connaissances dans le contexte de l'éducation humaine.

Concernant DistilBERT, il est pré-entraîné sur le même corpus que BERT, à savoir BookCorpus et English Wikipedia. Sa structure modèle est essentiellement similaire à celle de BERT, mais elle est simplifiée au niveau de la partie encodeur. Elle conserve un mécanisme d'auto-attention multicouches et un réseau neuronal *feed-forward*, mais le nombre de couches et de têtes d'attention est réduit à 6, nombre de paramètres est réduit à 66 millions.

Malgré sa réduction à seulement 60 % des paramètres de BERT, DistilBERT parvient à atteindre des performances comparables, voire supérieures, dans de nombreuses tâches de traitement du langage naturel. Dans le cadre du référentiel GLUE, il conserve 97 % des performances globales de BERT, tout en s'entraînant 60 % plus rapidement. La réduction du nombre de paramètres permet à DistilBERT d'afficher une taille de modèle de seulement 207 Mo, renforçant ainsi sa capacité à être déployé dans des environnements à ressources limitées tels que les appareils mobiles et les systèmes embarqués.

2.4.4. Tutoriel de DistilBERT

Sur le blog officiel de DistilBERT (Sanh et al., 2020) hébergé sur le site d'Hugging Face, de nombreux tutoriels officiels sont proposés pour différentes tâches, notamment pour la

tokenisation, les systèmes de question-réponse, la classification de textes et la classification de tokens.

La tokenisation est un processus de découpage d'un texte en unités linguistiques plus petites, appelées "tokens". Les tokens peuvent être des mots, des caractères ou des sous-mots ou d'autres unités significatives. La tokenisation est souvent la première étape dans le traitement du langage naturel, car elle prépare le texte pour des analyses plus approfondies.

Les systèmes de question-réponse consistent à comprendre une question formulée en langage naturel et à fournir une réponse pertinente à partir d'une source de données. Quant à la classification de textes, l'objectif est de catégoriser un texte donné dans l'une des catégories prédéfinies. Par exemple, classer des commentaires en ligne comme positifs, négatifs ou neutres.

Finalement, étant donné que nous nous intéressons à trouver les oronymes dans les textes, notre mission s'inscrit donc dans la dernière catégorie mentionnée ci-dessus, la *classification de tokens*. Cette tâche vise à identifier et classer des entités dans un texte avec des catégories sémantiques spécifiques, telles que des noms de personnes, de lieux, d'organisations, de dates, d'heures, etc. À cette fin, le texte d'entrée est segmenté en tokens individuels, et puis, une catégorie ou une étiquette spécifique est prédite pour les tokens qui le composent.

Dans le tutoriel[24] sur la classification de tokens, le processus est présenté étape par étape en commençant par le chargement de jeux de données, le pré-traitement des jeux de données et enfin le réglage fin (*fine-tuning*) du modèle. Nous expliquerons chacune de ces étapes plus en détails dans le chapitre suivant.

[24]Lien vers le notebook : https://colab.research.google.com/github/huggingface/notebooks/blob/main/examples/token_classification-tf.ipynb

Chapitre 2. Méthodologies

Ce chapitre propose une exploration approfondie des étapes clés impliquées dans la mise en œuvre de ce stage, avec une attention particulière portée aux approches adoptées pour réaliser les objectifs préalablement définis. La première étape englobe une présentation détaillée du processus de prétraitement des données du corpus, mettant en exergue les méthodologies employées pour organiser les données en vue d'une exploitation optimale. La séquence suivante du stage se focalise sur l'entraînement des modèles. Nous exposerons ici les stratégies déployées pour améliorer les performances des modèles pré-entraînés choisis.

Par la suite nous nous concentrerons sur les expériences menées avec les différents modèles pré-entraînés, en mettant l'accent sur les détails des tests et les ajustements effectués pour améliorer les résultats. La comparaison des différentes méthodologies nous aidera à évaluer leurs performances individuelles et à sélectionner l'approche la plus appropriée. En somme, ce chapitre fournira un aperçu approfondi des actions concrètes menées à chaque étape de la période de stage, offrant ainsi un aperçu détaillé du déroulement de nos travaux.

1. Préparation du jeu de données

Durant cette étape, nous effectuons un prétraitement à la fois sur le corpus et sur la liste des oronymes mentionnés précédemment. Cette opération est réalisée en élaborant des scripts Python[25] distincts pour chaque tâche, qui sont ensuite intégrés au sein d'une classe Python globale. L'objectif est de parvenir à la création de l'ensemble de données requis pour l'entraînement du modèle en une seule exécution. Les étapes spécifiques seront minutieusement détaillées dans les sections suivantes.

1.1. Prétraitement des oronymes

Précédemment, nous avons mentionné une liste d'oronymes obtenus manuellement. Dans une première phase, cette liste, comprenant au total 1221 entrées, dont 1148 avec le déterminant anglais "*the*" et 73 sans ce dernier, a été traitée dans le but de préparer un jeu de données pour l'entraînement du modèle. Nous avons choisi d'adopter une approche consistant à les incorporer dans une expression régulière.

1.1.1. La notion d'expression régulière

[25] Source : https://www.python.org/

Une expression régulière (*regular expression*), également connue sous le nom de "regex", est composée de différents éléments, tels que des littéraux (représentant des caractères spécifiques), des métacaractères (symboles spéciaux ayant une signification particulière) et des quantificateurs (spécifiant le nombre d'occurrences d'un élément). Elle nous confère la capacité d'effectuer des recherches puissantes et des traitements avancés sur des données textuelles. Parmi ses nombreuses applications, on compte la validation de formats, la recherche de motifs spécifiques, le remplacement de texte et l'extraction d'informations. Les expressions régulières sont ainsi un outil clé pour assurer un traitement efficace et précis des données textuelles.

L'origine de cette notion remonte aux années 1950, lorsque le mathématicien américain Stephen Kleene l'a introduite en tant que formalisme pour décrire les langages formels. Par la suite, son utilité s'est étendue au traitement de texte et aux opérations de recherche. Les années 1960 et 1970 ont vu une montée en popularité grâce au travail de Ken Thompson, qui a intégré les expressions régulières dans les outils Unix[26], notamment dans la commande grep[27]. Cette démarche a joué un rôle majeur dans la diffusion et l'adoption des expressions régulières dans le domaine de la programmation et du traitement de texte. Depuis lors, de nombreux langages de programmation et environnements ont intégré les expressions régulières dans leurs fonctionnalités, faisant d'elles un élément indispensable au sein de nombreux outils logiciels destinés à la manipulation des données textuelles.

1.1.2. Compilation de motif des oronymes

Une grande majorité des oronymes que nous avons à notre disposition contiennent le déterminant anglais "the". Nous avons observé que cela se produit dans trois cas distincts : premièrement, lorsque l'instance spécifique se trouve en début de phrase et s'écrit "The" ; deuxièmement, lorsque les trois lettres sont en majuscules, c'est-à-dire "THE", bien que cela soit moins fréquent ; et enfin, le troisième cas, le plus courant, est lorsque "the" est utilisé au milieu d'une phrase. Ces observations ont conduit à une considération importante lors de la préparation de nos expressions régulières.

Afin de pouvoir faire correspondre simultanément les résultats avec et sans toutes les formes de "the" dans le corpus total, nous avons retiré ces formes, conservé uniquement le corps

[26] UNIX : Uniplexed Information and Computing Service, est un système d'exploitation informatique multi-utilisateurs et multiprocessus développé par Kenneth Thompson et Dennis Ritchie dans les années 1970 ; Source : https://www.opengroup.org/unix

[27] Grep : un outil de ligne de commande utilisé à l'origine sur le système d'exploitation Unix, écrit pour la première fois par Ken Thompson.

de chaque oronyme, les avons encadrés entre parenthèses et avons ajouté une barre verticale dans chacun d'entre eux, comme suit :

```
(Tête noire|Col de Balme|Glacier du Tour|…)
```

Par la suite, nous utilisons le motif `(?i)\b(the\s+)?()\b` pour identifier les syntagmes nominaux commençant par "the", en tenant compte de la correspondance insensible à la casse. Nous pouvons obtenir une expression régulière complète en combinant ces deux parties comme suit :

```
(?i)\b(the\s+)?(Tête noire|Col de Balme|…)\b
```

- `(?i)` indique que la correspondance doit être insensible à la casse dans l'expression régulière.
- `\b` marque une limite de mot pour garantir que le syntagme nominal correspondant est pris en compte dans un contexte de mot.
- `(the\s+)?` correspond à l'élément optionnel "the", suivi d'un ou plusieurs espaces (ou autres caractères vides), l'espace étant également facultatif.
- `(Tête noire|…)` regroupes plusieurs candidats séparés par la ligne verticale `|`. Il indique que n'importe laquelle de ces sous-expressions peut être trouvée dans le texte, ce qui signifie qu'elle peut correspondre à l'un de ces oronymes. Par exemple, elle peut correspondre à *Tête noire*, *Col de Balme*, ou *Glacier du Tour*.

Cela nous permet de capturer les différentes occurrences de ces oronymes dans notre corpus, y compris celles contenant des variations de "the" mentionnées précédemment. L'expression complète est enregistrée dans un fichier nommé "`pattern.txt`" en vue d'une utilisation ultérieure.

1.2. Prétraitement du corpus d'entraînement

Comme mentionné dans la section 2.2 du premier chapitre, le corpus se compose de 384 fichiers au format XML qui sont annotés selon la norme TEI 1.0. La structure principale de chaque article, constituée de paragraphes, est balisée avec `<p>` et `</p>`, tandis que d'autres éléments tels que le titre, le sous-titre et le numéro de page sont également balisés avec leurs balises respectives. Par exemple, `<title></title>` est utilisé pour baliser le titre, et `<fw type="pageNum">2</fw>` est utilisé pour baliser le numéro de page, représentant ainsi la deuxième page dans ce cas précis.

Cependant, le contenu le plus pertinent pour atteindre notre objectif est celui qui est encapsulé entre les balises et . Par conséquent, notre prochaine étape consiste à extraire ce contenu spécifique. À cette fin, nous utilisons à nouveau un script Python.

La méthodologie implique l'utilisation de la bibliothèque OS de Python pour repérer tous les fichiers avec l'extension .xml dans le dossier du projet. Ensuite, la bibliothèque BeautifulSoup[28] sera employée pour extraire tout le contenu commençant par une balise , fichier par fichier. Ensuite, nous utiliserons la fonction de tokenisation de la bibliothèque NLTK (Loper & Bird, 2002)[29] pour récupérer les contenus de l'étape précédente.

NLTK est développé depuis 2001 par Edward Loper et Steven Bird de l'université de Pennsylvanie, il s'agit d'une bibliothèque Python populaire à libre accès qui intègre une série d'outils et de ressources pour le traitement et l'analyse de données textuelles, y compris la désambiguïsation, l'annotation, l'analyse syntaxique, l'analyse sémantique, et bien plus encore.

Par la suite, nous enregistrons le contenu des balises ligne par ligne, fichier par fichier, en conservant le nom de fichier d'origine et au format .txt, dans le chemin de sortie spécifié. En procédant ainsi, nous prenons en compte d'éventuelles utilisations ultérieures. De plus, nous avons sauvegardé un fichier distinct nommé corpus.txt, qui englobe les contenus de toutes les balises et dans tous les fichiers XML, autrement dit la somme de tous les fichiers TXT individuels. Les deux fonctions qui mettent en œuvre les traitements ci-dessus sont `"load_xml_files"` et `"process_xml_files"`.

Nous utilisons ensuite l'expression régulière présentée dans la section précédente sur ce fichier du corpus, permettant ainsi l'identification des phrases du corpus contenant les oronymes spécifiés dans le fichier. À cette fin, nous avons conçu une fonction appelée `"process_corpus_file"`.

Pour commencer, nous établissons les chemins d'accès aux fichiers d'entrée et de sortie. Ces fichiers incluent `"corpus.txt"` ainsi que quatre fichiers de sortie distincts. Ces quatre fichiers ont pour rôle d'enregistrer différents types de correspondances.

- `"matched_occurrences.txt"` pour conserver les oronymes correspondants (voir la figure 4).

[28] Source : https://pypi.org/project/beautifulsoup4/
[29] NLTK : Natural Language Toolkit; Source : https://www.nltk.org/

- "matched_lines.txt" pour les lignes de texte avec des correspondances d'oronymes (*idem*).
- "matched_occurrences_and_lines.txt" pour stocker simultanément les deux éléments mentionnés ci-dessus(voir la figure 5).
- "list_without_oronymes.txt" pour sauvegarder les lignes de texte sans oronymes correspondants (voir la figure 6).

```
the Glacier de Trient
the Glacier de Salena
the Glacier du Tour
the Aiguille du Tour
the Glacier du Tour
the Mont Blanc
Mont Blanc
the Col de Balme
the Grand Plateau
The Glacier de Taconnay
The Glacier des Bossons
```

```
The Glacier du Tour keeps a general direction towards the north-west, while
that of Salena, which is more sinuous, is turned a little to the south of east.
It is not, however, so simple a matter as might be supposed to pass from the
Glacier du Tour to that of Salena.
The Glacier du Tour, the twin system of the Trient and Orny and the Glacier de
Salena, are all on different levels; the Glacier
du Tour being much the highest of the three; the head of the Trient occupying
an intermediate level, and the Glacier de Salena being much lower than either.
There is a difference of probably not less than 1000 or 1500 feet between the
level of the highest plateau of the Glacier du Tour and that of the portion of
the Glacier de Salena which lies immediately behind the rocky boundary
separating the two; and the precipitous nature of the southern face of the
dividing range (above the Glacier de Salena) forbids all thought of passing
directly across it.
```

Figure 4. Exemples d'occurrences d'oronymes (à gauche) et exemples de phrases contenant des oronymes (à droite).

```
The Glacier de Taconnay
The Glacier de Taconnay, almost as high as the Grand Mulets, was dusted over
with the dirt blown from the rocks.
The Glacier des Bossons
The Glacier des Bossons was dirtier than I ever saw it before, and when I
walked up after breakfast to Balmat's cottage, a few hundred yards above the
church at Chamouni, it was at times with difficulty that I kept my legs.
```

Figure 5. Occurrences d'oronymes et leurs contextes.

```
It was just eight o'clock when we bid adieu to the landlord, and left our
homely, but clean and hospitable, quarters for the
trackless waste of ice and snow which lay between us and the next human
habitation we should see.
We lingered here a few moments, and while doing so the mists cleared swiftly
away and disclosed to our wondering eyes a vast series of plateaus, swelling
domes, and steep banks of ice, stretching back from the point above which we
stood to the origin of the glacier, a distance of many miles.
```

Figure 6. Lignes de texte sans oronymes correspondants.

Le reste de la fonction "process_corpus_file" est présenté dans le pseudo-code ci-dessous :

```
1. Ouvrir fichier_pattern en mode lecture
2. Lire contenu de fichier_pattern
3. Compiler modèle_expression_régulière avec contenu
4.
5. occurrences_correspondantes = LISTE_VIDE
6. lignes_correspondantes = LISTE_VIDE
```

```
lignes_sans_oronymes = LISTE_VIDE

Ouvrir fichier_entrée en mode lecture
Lire texte de fichier_entrée

correspondance_trouvée = FAUX

Pour chaque ligne dans texte :
    Si modèle_expression_régulière correspond à ligne :
        correspondance_trouvée = VRAI
        Ajouter modèle_expression_régulière à occurrences_correspondantes
        Ajouter ligne à lignes_correspondantes
    Sinon :
        Ajouter ligne à lignes_sans_oronymes
    Fin Si
Fin Pour

Si correspondance_trouvée = FAUX :
    Retourner ''

Ouvrir fichier_sortie_1 en mode écriture
Écrire occurrences_correspondantes dans fichier_sortie_1

Ouvrir fichier_sortie_2 en mode écriture
Écrire lignes_correspondantes dans fichier_sortie_2

Ouvrir fichier_sortie_3 en mode écriture
Pour chaque occurrence, ligne dans occurrences_correspondantes Et
    lignes_correspondantes :
    Écrire occurrence Et ligne dans fichier_sortie_3
Fin Pour

Ouvrir fichier_sortie_4 en mode écriture
Écrire lignes_sans_oronymes dans fichier_sortie_4
```

En résumé, ce script utilise des expressions régulières pour identifier les oronymes et enregistre les résultats dans différents fichiers de sortie. Cette approche permet de classer et d'organiser de manière structurée les informations pertinentes contenues dans le corpus.

1.3. Annotation

Après avoir accompli les étapes précédentes, nous avons extrait toutes les phrases du fichier "`corpus.txt`" qui contiennent un oronyme listé dans le motif, et les avons regroupées dans le fichier "`matched_lines.txt`". Cependant, il nous reste une étape cruciale à accomplir : l'annotation de ces phrases en vue de l'utilisation ultérieure, car pour un apprentissage supervisé, il faut entraîner un modèle sur du corpus déjà annoté.

1.3.1. Les étiquettes et la notion de BIO dans NER

Étant donné que notre tâche relève de la reconnaissance des entités nommées (NER), une démarche fréquente dans le domaine du traitement du langage naturel visant à identifier et classer les entités nommées telles que les noms de personnes, de lieux, d'organisations, etc.,

nous désignons ces trois types d'entités nommées susmentionnées par les étiquettes PER, LOC et ORG, respectivement. En outre, d'autres étiquettes couramment employées sont présentées dans le tableau ci-dessous :

Étiquettes	Objets
TIME	les expressions temporelles telles que les dates, les périodes, etc
DATE	une date spécifique
MONEY	les montants monétaires
PERCENTAGE	les valeurs en pourcentage
ORDINAL	les mots ordinaux, tels que "premier", "deuxième", etc
QUANTITY	les mots de quantité, tels que "trois livres", "cinq kilogrammes", etc
EVENT	les événements spécifiques
PRODUCT	les produits ou des marchandises
WORK_OF_ART	les œuvres artistiques, de la littérature, etc.
NORP	le nom d'un pays, d'une nation, etc
LAW	les noms de lois, de règlements, d'articles juridiques
MISC	une catégorie de regroupement pour diverses entités nommées qui ne sont pas facilement classifiables

Tableau 2. Étiquettes usuelles pour la reconnaissance des entités nommées (NER)

Ces étiquettes jouent un rôle essentiel en caractérisant et en classifiant diverses entités nommées présentes dans le texte, ce qui simplifie grandement leur identification et leur traitement subséquent au sein d'une tâche de reconnaissance des entités nommées. Elles offrent la flexibilité d'être adaptées et étendues en fonction des exigences spécifiques des tâches et des jeux de données, permettant ainsi de répondre aux besoins d'identification propres à différents domaines et types d'entités. La sélection et la définition minutieuses de ces étiquettes revêtent une importance capitale pour une identification et une classification précises des entités nommées.

Notre mission se concentre exclusivement sur l'identification des oronymes, qui appartiennent à la catégorie des noms de lieux, ce qui nous amène à sélectionner l'étiquette LOC pour cette identification.

Par la suite, il est essentiel de comprendre une méthodologie répondue concept dans la reconnaissance des entités nommées (NER), à savoir le format BIO (*Beginning-Inside-Outside*) (Ramshaw & Marcus, 1995). L'acronyme BIO fait référence à l'intérieur, à l'extérieur et au début d'une entité, respectivement. Le format BIO constitue un schéma d'étiquetage largement utilisé pour annoter les entités nommées au sein d'un texte. Cette méthode a été proposée par Lance A. Ramshaw et Mitchell P. Marcus en 1995 dans leur article *Text Chunking using Transformation-Based Learning*.

En ce qui concerne notre tâche actuelle portant sur l'identification des noms de lieux (LOC, *Location*), nous mettons en œuvre la méthode suivante :

B-LOC (*Beginning-Location*) : Cette étiquette indique le commencement d'une entité toponymique. Elle est utilisée pour signaler le début d'une nouvelle entité toponymique.

I-LOC (*Inside-Location*) : Cette étiquette indique la partie interne d'une entité toponymique. Lorsqu'une entité toponymique est constituée de plusieurs tokens, tous sauf le premier sont étiquetés avec "I-LOC".

O (*Outside*) : Cette étiquette est utilisée pour marquer la partie du texte qui ne fait pas partie d'une entité toponymique, c'est-à-dire la portion de texte extérieure à toute entité toponymique. Prenons l'exemple d'une phrase tirée de notre corpus :

Phrase d'origine :

"*They had reached the Faulberg about eight, and slept more soundly than on the previous night.*[30]"

Après l'étiquetage BIO :

```
"They[O] had[O] reached[O] the[B-LOC] Faulberg[I-LOC] about[B-LOC] eight[O] ,[O]
and[O] slept[O] more[O] soundly[O] than[O] on[O] the[O] previous[O] night[O] .[O]"
```

Dans cette phrase, nous pouvons illustrer l'application du format BIO pour marquer l'entité toponymique *the Faulberg*. L'étiquette "B-LOC" est attribuée à "the", signalant ainsi le début de cet oronyme, suivi de "I-LOC" pour *Faulberg*, dénotant la partie intérieure. Les tokens qui ne font pas partie de l'entité reçoivent l'étiquette "O" (extérieur).

Ce système d'étiquetage joue un rôle essentiel dans l'identification et l'extraction des entités nommées dans le texte, en permettant de distinguer les différentes parties de ces entités. En combinant le schéma d'étiquetage BIO, la tâche NER peut capturer avec une précision

[30] Traduction : Ils avaient atteint le Faulberg vers huit heures et avaient dormi plus profondément que la nuit précédente.

accrue les informations relatives aux entités nommées, notamment les noms géographiques. Cela conduit à une représentation sémantique plus riche, qui favorise une analyse et un traitement plus approfondis du texte. Cette démarche permet une détection précise et une classification efficace des noms de lieux dans le texte.

1.3.2. Mise en œuvre de l'annotation

Cependant, la mise en pratique de ce qui a été décrit précédemment nécessite quelques ajustements pour assurer la reconnaissance précise du format BIO par le modèle DistilBERT. Il ne suffit pas uniquement de baliser les phrases dans le fichier matched_lines.txt comme illustré dans l'exemple précédent. Le modèle doit être capable de comprendre le schéma d'étiquetage. Plutôt que d'utiliser les étiquettes B-LOC, I-LOC et O telles quelles, nous adoptons une approche numérique en utilisant les valeurs 0, 1 et 2. Dans ce schéma, le chiffre 0 correspond à la partie initiale d'une entité nommée (équivalent à B-LOC), le chiffre 1 correspond à la partie intérieure (équivalent à I-LOC), tandis que le chiffre 2 désigne la partie en dehors d'une entité nommée (équivalent à O). Cette transformation permet au modèle DistilBERT de mieux saisir le schéma d'étiquetage lors de l'entraînement et de la prédiction.

Nous avons développé une fonction essentielle appelée "`process_matched_lines`" pour effectuer cette tâche spécifique. Son rôle est de traiter les lignes de texte correspondantes, d'attribuer les étiquettes NER appropriées à chaque token et d'enregistrer les résultats dans le fichier de sortie. Voici un script en pseudo-code démontrant son fonctionnement :

```
1. fonction process_matched_lines(matched_lines, matched_occurrences,
   output_file):
2.     Ouvrir le fichier de sortie en mode écriture au format jsonlines
3.
4.     Pour chaque index i, chaque ligne dans matched_lines :
5.         Diviser la ligne en tokens
6.
7.         Créer une liste ner_tags initialisés avec '2' pour chaque token
8.
9.         Diviser les occurrences correspondantes en mots individuels
10.        occurrences = diviser matched_occurrences[i] en utilisant la virgule
           comme séparateur
11.
12.        Pour chaque occurrence dans occurrences :
13.            Initialiser l'index de début start_index à 0
14.            Tant que vrai :
15.                Essayer :
16.                    index = trouver l'indice du premier mot de l'occurrence.
                        dans la liste de tokens à partir de start_index
17.                    Si les tokens[index:index + longueur de l'occurrence]
                        sont égaux à l'occurrence:
```

```
                        Pour chaque j, chaque token_index dans la plage
                        d'indice de index à index + longueur de l'occurrence :
                            Si j est égal à 0 :
                                Définir ner_tags[token_index] à '0'
                            Sinon :
                                Définir ner_tags[token_index] à '1'
                        casser la boucle
                    Sinon :
                        Définir start_index à index + 1
                    Fin Si
                attraper une ValueError:
                    Sortir de la boucle
            Fin Tant que
        Fin Pour

        Écrire les résultats dans le fichier de sortie au format jsonlines :
        ({
            'id': 'nominalp#' suivi de l'index de ligne de texte,
            'tokens': tokens,
            'ner_tags': ner_tags
        })
    Fin Pour
```

Enfin, dans le fichier de sortie appelé "`output.jsonl`", nous avons obtenu 11 222 résultats. Ils sont tous étiquetés comme dans l'exemple ci-dessous :

`{"id": "nominalp#00085", "tokens": ["As", "constantly", "happens", "in", "the", "Alps", ",", "this", "heat", "was", "the", "precursor", "of", "rain", "."], "ner_tags": ["2", "2", "2", "2", "0", "1", "2", "2", "2", "2", "2", "2", "2", "2"]}`

Nous avons choisi le format JSONL (JSON Lines)[31] parce qu'il permet de stocker des données structurées dans un format texte. Il est similaire à JSON[32], mais chaque ligne représente un objet JSON distinct au lieu de placer tous les objets dans un grand tableau JSON. Le format JSONL est souvent utilisé pour stocker un grand nombre d'enregistrements de données, chaque ligne représentant un seul enregistrement, ce qui, dans certains cas, est plus efficace que de contenir tous les enregistrements dans un grand tableau.

Cette dernière étape a pour objectif de simplifier les futures étapes de traitement et d'analyse. En somme, cette fonction revêt une importance capitale en préparant les données pour que le modèle DistilBERT puisse interpréter correctement le format BIO. Elle attribue les

[31] Source : https://jsonlines.org/

[32] JSON : JavaScript Object Notation, initialement conçu par le programmeur américain Douglas Crockford dans les années 2000, est un format léger d'échange de données couramment utilisé pour représenter des données structurées. Il offre une facilité de lecture et d'analyse, et convient à une variété de langages de programmation; Source : https://www.json.org/json-en.html

étiquettes NER appropriées à chaque token du texte et enregistre ces informations dans les fichiers de sortie correspondants.

1.4. Division du « split »

Dans cette étape, qui est également la dernière étape de la phase de préparation de notre jeu de données. Nous devons diviser le fichier "`output.jsonl`" obtenu à l'étape précédente en trois sous-ensembles d'entraînement distincts. Ces ensembles, nommés "train", "validation" et "test", sont essentiels pour le processus d'apprentissage automatique. Cette méthode bien établie nous permet d'entraîner, d'ajuster et d'évaluer nos modèles de manière efficace.

1.4.1. L'ensemble "train"

Parmi ces ensembles, "train" est dédié à l'entraînement de notre modèle, et il contient 7 855 phrases, représentant 70% de l'ensemble des données. Cette proportion significative est choisie pour favoriser la généralisation du modèle, lui permettant de s'adapter à diverses situations et variations. En effet, un ensemble d'apprentissage plus vaste contribue à mieux saisir les caractéristiques sous-jacentes des données, assurant que notre modèle intègre de manière approfondie les éléments essentiels des données. Cette approche se traduira par une performance optimisée lors des évaluations futures.

L'ensemble "train" constitue la base de l'apprentissage, exposant notre modèle à une variété de cas et de schémas. En le nourrissant de ces données à plusieurs reprises, le modèle acquiert la capacité d'identifier les relations complexes entre les données, ce qui le prépare à effectuer des prédictions précises sur de nouvelles données.

1.4.2. L'ensemble "validation"

L'ensemble "validation", quant à lui, est dédié à l'ajustement des hyperparamètres du modèle, tels que le taux d'apprentissage (*learning rate*)[33] et la régularisation (regularization)[34]. Pendant la phase d'entraînement, les performances obtenues sur l'ensemble de validation sont

[33] Taux d'apprentissage : un paramètre qui détermine la taille des pas que l'algorithme de descente de gradient effectue lors de la mise à jour des poids du modèle. Un taux d'apprentissage plus élevé permet des mises à jour plus importantes, ce qui peut accélérer la convergence, mais peut également entraîner des sauts excessifs et une convergence instable. Un taux d'apprentissage plus faible peut favoriser une convergence plus stable, mais peut ralentir l'entraînement. Le choix optimal du taux d'apprentissage est crucial pour obtenir une formation efficace et rapide.

[34] Régularisation : sert à prévenir le surapprentissage (overfitting) d'un modèle. Lorsque les modèles sont trop complexes, ils peuvent mémoriser les données d'entraînement au lieu d'apprendre des modèles généraux. Cela conduit à une performance médiocre sur de nouvelles données. La régularisation introduit des termes supplémentaires dans la fonction de perte qui pénalisent les valeurs élevées des poids du modèle. Cela encourage le modèle à généraliser mieux en évitant des valeurs de poids excessives. Des techniques courantes de régularisation comprennent la régularisation L1 (Lasso) et L2 (Ridge), qui ajoutent des termes de pénalité basés sur les valeurs absolues ou au carré des poids, respectivement.

utilisées pour sélectionner les paramètres optimaux et les architectures de modèle les plus performantes. La présence d'un ensemble de validation empêche le modèle de surapprendre les données d'entraînement. Lorsque le surapprentissage (*overfitting*) se produit, le modèle peut capturer des relations qui ne sont spécifiques qu'aux données d'entraînement et ne s'appliquent pas de manière générale à d'autres données similaires. Cela peut se manifester par une performance médiocre lors de la prédiction de nouvelles données, car le modèle a perdu sa capacité à discerner les caractéristiques réellement importantes des données. Pour éviter le surapprentissage, plusieurs techniques peuvent être utilisées, telles que :

- L'utilisation de données variées pour exposer le modèle à diverses situations et améliorer sa capacité de généralisation.
- Diviser les données en ensembles d'entraînement distincts permet de surveiller les performances sur des données non vues et d'ajuster en conséquence.
- Les techniques de régularisation, comme L1 et L2, ajoutent des termes de pénalité pour limiter les poids excessifs.
- Réduire la complexité du modèle en ajustant le nombre de couches ou de neurones est également utile.
- Le *dropout* : cela consiste à désactiver aléatoirement des neurones pendant l'entraînement pour éviter une dépendance excessive.
- L'*Early Stopping* : surveille les performances sur l'ensemble de validation et arrête l'entraînement lorsque la performance se détériore, empêchant un ajustement excessif aux données d'entraînement.

L'objectif est d'obtenir un équilibre entre une bonne performance sur les données d'entraînement et la capacité du modèle à généraliser correctement sur de nouvelles données.

Nous avons choisi de consacrer 2 244 phrases, soit 20 % de nos données à l'ensemble de validation. Cette taille plus réduite accélère le processus de développement du modèle en permettant une évaluation rapide des performances pour différentes combinaisons de paramètres et d'architectures. De plus, l'utilisation d'un ensemble de validation indépendant de l'ensemble d'entraînement et de l'ensemble de test prévient toute fuite d'informations lors de l'ajustement des paramètres, garantissant ainsi l'intégrité de l'évaluation et du développement du modèle.

1.4.3. L'ensemble "test"

L'ensemble "test" est réservé à l'évaluation des performances d'un modèle entraîné, simulant ainsi son comportement dans des scénarios réels. Cet ensemble est constitué de données non exposées au modèle pendant les phases d'entraînement et de validation. L'évaluation sur l'ensemble de test fournit une évaluation plus précise de la capacité de généralisation du modèle et de son impact réel. Cela permet de déterminer si le modèle présente une sur-adaptation ou une sous-adaptation aux données. Notre ensemble de test contient 1 123 phrases, représentant 10 % de l'ensemble des données. Cette proportion offre suffisamment de données pour évaluer les performances du modèle sur des données inédites, fournissant ainsi des informations cruciales sur sa capacité à généraliser et à produire des résultats concrets dans des situations réelles.

La division de l'ensemble de données en trois sous-ensembles vise principalement à garantir la fiabilité et la capacité de généralisation du modèle. Cette stratégie permet de régler les paramètres et l'architecture du modèle pour éviter le surapprentissage tout en fournissant une évaluation impartiale de ses performances. Elle nous guide dans la prise de décisions plus éclairées lors du développement du modèle, aboutissant à des améliorations plus notables.

1.4.4. Mise en œuvre de la division

Nous commençons par importer les modules essentiels, "`json`" pour traiter les données au format JSONL et "`sklearn.model_selection`" pour diviser l'ensemble de données en sous-ensemble. Nous définissons ensuite une fonction appelée "`split_save_data`" qui accepte le chemin du fichier d'entrée, celui du fichier de sortie, les ratios des trois sous-ensembles et "`random_state`" (pour diviser aléatoirement l'ensemble des données) comme argument. À l'intérieur de la fonction, nous lisons le fichier JSONL d'entrée, analysons chaque ligne de données pour en faire un objet dictionnaire, et formons une liste de ces dictionnaires que nous stockons dans la variable "`data`".

Ensuite, nous utilisons la fonction "train_test_split" pour répartir aléatoirement l'ensemble de données en ensembles d'entraînement, de validation et de test. Ce processus de répartition se fait de manière aléatoire, en fonction d'un ratio de 70% pour l'ensemble d'apprentissage, 20% pour l'ensemble de validation et 10% pour l’ensemble de test.

À partir de ces ensembles de données divisés, nous inscrivons les données de l'ensemble d'entraînement, ligne par ligne, dans un fichier de sortie nommé "`train.jsonl`", les données de l'ensemble de validation dans le fichier "`validation.jsonl`" et les données de l'ensemble de test dans le fichier "`test.jsonl`".

Par la suite, nous affichons la taille de chaque sous-ensemble, c'est-à-dire le nombre d'échantillons dans les ensembles d'apprentissage, de validation et de test.

Pour poursuivre, nous définissons le chemin du fichier d'entrée, le chemin du répertoire de sortie et les taux de répartition.

Enfin, nous appelons la fonction `"split_save_dataset"` pour enregistrer les fichiers divisés dans le répertoire de sortie spécifié au format JSONL.

Après avoir achevé toutes les étapes mentionnées précédemment, incluant la préparation de la liste des oronymes, la création d'une expression régulière pour localiser les oronymes dans le corpus, la mise en forme du corpus (filtrage et conversion des contenus de paragraphes des documents XML en format texte TXT, repérage des correspondances des oronymes dans le texte à l'aide de l'expression régulière et enregistrement des résultats), ainsi que l'étiquetage des résultats selon le schéma BIO, et enfin la subdivision en trois sous-ensembles distincts, nous parvenons obtenir un ensemble de données prêt à être utilisé pour l'entraînement du modèle. Cette phase de préparation des données marque la clôture de la première étape de notre démarche, et tous les codes utilisés sont disponibles en annexe.

2. Entraînement du modèle DistilBERT

Dans cette étape, nous entrons dans la phase cruciale de l'entraînement du modèle. Nous mobilisons un ensemble de données méticuleusement préparé pour enseigner au modèle à identifier avec précision les entités nommées géographiques. Cette phase revêt une importance cruciale dans la création et la formation de notre modèle de traitement du langage naturel (TAL). Au préalable, nous avons exposé les caractéristiques des trois modèles, à savoir Transformer, BERT et DistilBERT, ainsi que leurs interrelations, dans le premier chapitre. Pour cette tâche spécifique, nous optons pour DistilBERT-base (ci-après DistilBERT) en tant que modèle de base et cherchons à rehausser ses performances dans la reconnaissance des oronymes en affinant les paramètres et en optimisant le processus d'entraînement.

Au chapitre 1, nous avons brièvement mentionné la disponibilité de nombreux tutoriels officiels pour diverses tâches sur la page officielle de DistilBERT sur Hugging Face. Notre tâche actuelle se rapporte à la version pré-entraînée pour la classification des tokens.

En suivant le tutoriel (l'url vers le *notebook* se trouve dans la note de bas de page n° 21), nous examinerons en détails le processus complet d'entraînement du modèle. Cela comprendra l'étape de chargement des données, la construction du modèle, le choix de la fonction de perte

et la configuration de l'optimiseur. En appliquant de multiples itérations d'entraînement sur les données d'apprentissage conformément au guide, notre but est de guider le modèle à apprendre à identifier et à étiqueter de manière précise les oronymes à travers l'ensemble de notre corpus. Cette démarche vise à habiliter le modèle à effectuer une reconnaissance fiable des occurrences d'oronymes dans des textes nouveaux.

L'intégralité du processus d'entraînement a été réalisée sur la plateforme en ligne gratuite de Google, Google Colab (Colaboratory)[35]. Cette plateforme s'appuie sur l'environnement Jupyter Notebook[36] et permet aux utilisateurs de coder et d'exécuter du code Python directement dans leur navigateur, tout en créant des documents combinant du code, du texte et des images. Elle facilite également le partage de ces documents pour la collaboration.

Le choix d'utiliser Google Colab (ci-après Colab) présente plusieurs avantages pour notre tâche :

- Ressources de calcul gratuites : Colab met à disposition des ressources gratuites de GPU et de TPU (Jouppi et al., 2017)[37], ce qui est très bénéfique pour l'entraînement de modèles. En particulier, dans les tâches d'apprentissage en profondeur, l'utilisation de GPU ou de TPU peut accélérer considérablement le processus d'entraînement, ce qui permet de gagner du temps.
- Environnement en ligne : étant donné que notre tâche implique de vastes données textuelles et de l'entraînement de modèle, notre ordinateur local pourrait ne pas suffire en termes de ressources de calcul. En utilisant Colab dans un environnement en ligne, nous pouvons éviter les limitations en ressources de calcul, assurant ainsi l'efficacité de notre tâche.
- Bibliothèques et dépendances préinstallées : Colab préinstalle de nombreuses bibliothèques Python couramment utilisées et leurs dépendances, évitant ainsi la nécessité de passer du temps sur la configuration de l'environnement et permettant de se concentrer directement sur la tâche elle-même.

[35] Source : https://colab.research.google.com/
[36] Source : Jupyter Notebook est un environnement de calcul interactif et d'analyse de données open source, créé en 2014 par les scientifiques en informatique Fernando Pérez et Brian Granger. Il offre une interface interactive où les utilisateurs peuvent écrire et exécuter du code, tout en ayant la possibilité de créer des documents riches en contenu, comprenant du code, des graphiques, du texte et des éléments multimédias; https://jupyter.org/
[37] TPU : Tensor Processing Unit, est une unité de traitement spécialisée développée par Google en 2016 pour accélérer les tâches d'apprentissage en profondeur et de calcul intensif en utilisant la puissance du calcul parallèle.

- Pas besoin d'installation ni de configuration : Colab est une plateforme basée sur le navigateur, ce qui signifie qu'aucune installation de logiciel ni configuration d'environnement ne sont nécessaires. Cela permet de commencer rapidement et facilement à travailler.

- Facilité de partage et de collaboration : Colab repose sur Jupyter Notebook, permettant d'intégrer code, texte et images dans un même document, et facilitant le partage et la collaboration avec d'autres personnes pour la consultation et le travail en équipe. Cela est particulièrement utile dans les contextes de collaboration d'équipe ou de présentation des résultats.

En résumé, le choix de Colab vise à tirer pleinement parti de ses ressources de calcul et de son environnement en ligne, ainsi qu'à bénéficier de ses fonctionnalités de partage et de collaboration pratiques, afin de mener à bien l'entraînement des modèles et les expérimentations de manière plus efficace.

2.1. Chargement des données

2.1.1. Installation des bibliothèques

Tout d'abord, une série de préparations et de configurations sont effectuées avant le début de l'apprentissage du modèle. Les étapes sont les suivantes :

- Installation des bibliothèques requises : plusieurs bibliothèques Python nécessaires ont été installées à l'aide de la commande `"! pip install"` pour installer plusieurs bibliothèques Python nécessaires, notamment `"transformers"` pour la manipulation des modèles TAL, `"datasets"` pour accéder aux ensembles de données, `"seqeval`[38]`"` pour l'évaluation des séquences et `"huggingface_hub"` pour la gestion des modèles et des données.

- Connexion au compte Hugging Face : en important et en appelant les fonctions de la bibliothèque "huggingface_hub", nous pouvons nous connecter au compte Hugging Face pour partager et gérer les modèles et les données dans l'environnement Colab.

[38] Seqeval : est un framework Python pour l'évaluation de l'étiquetage des séquences. il permet d'évaluer la performance des tâches de chunking telles que la reconnaissance des entités nommées, l'étiquetage de la partie du discours, l'étiquetage des rôles sémantiques, etc; Source : https://pypi.org/project/seqeval/0.0.10/

- Installation de "git-lfs" : Avec la commande "`! apt install git-lfs`", nous pouvons installer l'extension Git[39] "git-lfs", qui est utilisée pour gérer les fichiers volumineux, en particulier les fichiers binaires volumineux tels que les modèles.
- Configurer les informations Git : Avec la commande "`! git config`", nous pouvons configurer une adresse électronique et un nom d'utilisateur Git global pour l'authentification lors de l'utilisation de Git pour le partage de modèles et d'autres opérations.
- Importation de bibliothèques et impression des versions : L'importation de la bibliothèque "`transformers`" et l'exportation des informations relatives à sa version afin de garantir une installation correcte et une correspondance des versions.
- Envoi de traces de données : la fonction "`send_example_telemetry`" envoie des traces de données de l'utilisation actuelle à Hugging Face pour aider à améliorer la fonctionnalité et la performance de la bibliothèque.

Ces étapes ont pour but d'assurer une configuration correcte de l'environnement, l'installation des bibliothèques requises et la connexion à Hugging Face pour l'entraînement et le partage ultérieur du modèle.

2.1.2. Chargements

Dans le contexte du chargement de données, nous définissons tout d'abord la variable "`task`" qui indique la nature de la tâche, puis nous spécifions le modèle pré-entraîné à utiliser, "`distilbert-base-uncased`", sa caractéristique "*uncased*" signifie qu'il ne fait pas de distinction entre les lettres majuscules et minuscules lors du traitement du texte, ce qui est imortant pour notre tâche puisque les entités géographiques peuvent apparaître sous différentes formes de casse. En plus, le modèle "`distilbert-base-uncased`" est relativement petit, avec moins de paramètres, ce qui le rend moins gourmand en ressources computationnelles et en temps par rapport à des modèles plus grands tels que BERT. Cela le rend adapté aux environnements avec des ressources limitées pour des expérimentations et développements. En outre, cette caractéristique aide à réduire le risque de surapprentissage, permettant au modèle de mieux généraliser sur de nouvelles données.

[39] Git : un système de contrôle de version distribué utilisé pour suivre et gérer les changements du code logiciel, permettant la collaboration entre plusieurs personnes, etc. Source : https://git-scm.com/

Par conséquent, le choix de "distilbert-base-uncased" comme modèle de base pour notre tâche découle d'une considération globale des compétences du modèle, des besoins en ressources de calcul et du support disponible.

Ensuite, nous fixons la taille des lots (*batch size*) à 16, ce qui correspond au nombre d'exemples traités simultanément lors de l'apprentissage. Cette pratique du traitement par lots est fréquemment employée dans l'apprentissage profond pour optimiser l'efficacité du calcul parallèle. La sélection d'une taille de lot appropriée est cruciale, car elle influe sur la vitesse et la stabilité de l'entraînement.

Dans le tutoriel, le choix d'une taille de lot de 16 peut résulter d'une évaluation pondérée des facteurs, notamment la nature du modèle, les ressources matérielles disponibles et la dimension de l'ensemble de données. Une taille de lot plus grande maximise généralement l'utilisation de la parallélisation, accélérant ainsi l'entraînement. Toutefois, une taille excessive peut entraîner des problèmes de mémoire et de capacité graphique, perturbant la stabilité. Une taille de lot plus réduite économise la mémoire, adaptée aux ressources limitées, mais peut légèrement ralentir l'apprentissage. Ce choix reflète un compromis équilibré entre ces paramètres pour obtenir des performances optimales.

Ensuite, nous allons présenter les étapes de chargement, de préparation et de visualisation de l'ensemble de données en suivant les étapes ci-dessous :

Tout d'abord, nous importons les fonctions de chargement de jeux de données et de mesures, à savoir "load_dataset" et "load_metric", depuis la bibliothèque "datasets". Ensuite, nous montons Google Drive en utilisant la fonction de montage propre à Colab. Par la suite, nous spécifions le chemin du répertoire de données pour pouvoir accéder aux fichiers. En utilisant la fonction "load_dataset", nous chargeons le jeu de données au format JSONL. Cela crée un objet "datasets" contenant les données. Nous passons les chemins d'accès vers les fichiers contenant les données d'entraînement, de validation et de test via le paramètre "data_files".

Pour examiner les informations des ensembles de données chargés, nous imprimons simplement l'objet "datasets". Nous obtenons les résultats suivants :

```
DatasetDict({
        train: Dataset({
              features: ['id', 'tokens', 'ner_tags'],
```

```
        num_rows: 7855 })
    validation: Dataset({
        features: ['id', 'tokens', 'ner_tags'],
        num_rows: 2244 })
    test: Dataset({
        features: ['id', 'tokens', 'ner_tags'],
        num_rows: 1123 })
})
```

Nous pouvons constater que chaque ensemble de données a des caractéristiques similaires, à savoir les colonnes "`id`", "`tokens`" et "`ner_tags`".

Ensuite, pour mieux comprendre la structure et le contenu des données, nous pouvons afficher le premier échantillon de l'ensemble d'apprentissage en utilisant "`datasets["train"][0]`".

Par la suite, nous définissons une fonction nommée "`show_random_elements`" pour présenter des échantillons de données choisis au hasard. Pour cela, nous importons les fonctions "`ClassLabel`" et "`Sequence`" de la bibliothèque "`datasets`". Ces fonctions servent à gérer les étiquettes de classe et les séquences de données. Ensuite, nous importons les modules "`random`", "`pandas`", ainsi que les fonctions "`display`" et "`HTML`" de la bibliothèque "`IPython.display`".

À l'intérieur de la boucle, un certain nombre d'échantillons choisis au hasard sont extraits et enregistrés dans un "`DataFrame`". Ce dernier est ensuite transformé en une forme plus lisible en fonction des types de caractéristiques présentes dans l'ensemble de données. Enfin, le contenu du DataFrame est affiché en utilisant la fonction "display(HTML)".

En dernier lieu, nous appelons la fonction "`show_random_elements`" pour afficher un certain nombre d'échantillons sélectionnés au hasard depuis l'ensemble d'apprentissage de manière interactive, comme représenté dans la figure 7.

```
[ ] show_random_elements(datasets["train"])
```

	id	tokens	ner_tags
0	nominalp#07062	[Aiguille, de, la, Za.—On, August, 81, ,, Mr., J., C., Leman, ,, accom¬, panied, by, the, guides, Frederic, Payot, and, Adolphe, Folignet, (, both, of, Chamonix,), ,, in, ascending, the, Aiguille, de, la, Za, from, Arolla, ,, took, a, new, route, to, the, base, of, the, peak, .]	[2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]
1	nominalp#06641	[Already, the, Alps, are, being, regarded, more, and, more, with, an, eye, rather, to, successful, work, than, to, general, enjoyment, .]	[2, 0, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]
2	nominalp#05025	[Paul, Giissfeldt, ,, of, Berlin, ,, with, Hans, Grass, ,, Peter, Jenni, ,, and, a, porter, ,, left, the, Misauna, Alp, in, the, Roseg, Thai, on, Sept., 12, ,, to, try, and, reach, the, very, well-marked, col, at, the, head, of, the, Tschierva, glacier, ,, between, Piz, Bernina, and, Piz, Roseg, .]	[2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 1, 2]
3	nominalp#02366	[The, route, to, the, Lauteraar, Sattel, is, ,, for, about, half, an, hour, after, leaving, the, sleeping-place, ,, the, same, as, that, to, the, Wetterhorn, ;, it, then, diverges, ,, turning, sharp, off, to, the, right, in, a, southerly, direction, .]	[2, 2, 2, 0, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]
4	nominalp#08329	[This, is, formed, by, the, junction, of, a, ridge, coming, from, the, south-west, *, (, on, which, is, the, Viso]	[2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 1]
5	nominalp#10324	[Bad, weather, continued, ,, so, I, left, ,, parting, with, my, good-, natured, guide, at, the, top, of, the, Pfitscher, Joch, .]	[2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 1, 1, 2]
6	nominalp#06920	[Within, these, boundaries, ,, there, are, three, main, ridges, ,, running, ,, roughly, speaking, ,, from, north, to, south, ,, the, highest, summits, of, which, (, reckoning, from, east, to, west,), are, the, Aiguilles, d, ', Arve, ,, the, Grandes, Rousses, ,, and, the, Belledonne, .]	[2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 1, 1, 2, 2, 2, 2, 2]
7	nominalp#02646	[Our, plan, was, to, take, up, our, quarters, for, the, night, as, near, as, possible, to, the, Silberhorn, Glacier, .]	[2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 1, 2, 2]
8	nominalp#05619	[Our, next, plan, was, ,, therefore, ,, to, explore, the, Floiten, Grund, ,, ascend, the, Loffel, ,, and, if, possible, to, descend, on, the, other, side, into, the, Ahren, Thai, ,, though, on, the, latter, point, Ball, gave, us, little, encouragement, .]	[2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]
9	nominalp#02308	[expanse, of, the, Sesia, glacier, lying, between, us, and, it, ,, though, from, our, position, it, was, difficult, to, form, any, estimate, of, the, distance, .]	[2, 2, 0, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]

Figure 7. Exemples de phrases de TRAIN SET affichées aléatoirement

En résumé, ces étapes détaillent le processus de chargement de l'ensemble de données, fournissent des informations essentielles sur les données, les étiquettes associées, et expliquent comment visualiser des exemples aléatoires. Cette phase constitue la fondation pour l'étape ultérieure de l'apprentissage du modèle et de l'analyse des données.

2.2. Prétraitement des données

Dans le processus de prétraitement des données, nous commençons par utiliser un Tokeniseur de Transformers. Ce tokeniseur divise les entrées en tokens, ce qui peut inclure la division des mots en sous-parties en utilisant l'algorithme BPE *(Byte Pair Encoding*) (Sennrich et al., 2016)[40] Ensuite, il convertit ces tokens en identifiants correspondants dans le vocabulaire pré-entraîné et les formate dans la structure attendue par le modèle. De plus, il génère les autres entrées nécessaires pour le modèle. Les étapes sont les suivantes :

Nous instancions d'abord notre tokeniseur à l'aide de la méthode `"AutoTokenizer.from_pretrained"`. Cela garantit que nous obtenons un tokeniseur correspondant à l'architecture du modèle que nous souhaitons utiliser et que nous téléchargeons le vocabulaire utilisé lors de la pré-entraînement de ce point de contrôle spécifique.

Nous nous assurons que notre tokeniseur est rapide et qu'il est de la bibliothèque `"Tokenizers"`. Ces tokeniseurs rapides sont disponibles pour presque tous les modèles et nous aurons besoin de certaines de leurs fonctionnalités spéciales pour notre prétraitement.

[40] BPE : est un algorithme de segmentation basé sur les sous-mots, qui décompose progressivement les mots du vocabulaire en parties plus petites, afin de mieux gérer les mots inconnus et la complexité linguistique.

Nous pouvons directement appeler ce tokeniseur sur une phrase, par exemple :

```
tokenizer("Bonjour, ceci est une phrase !")
```

Pour les entrées déjà divisées en mots, nous devons passer la liste des mots au tokeniseur avec l'argument "`is_split_into_words=True`", par exemple :

```
tokenizer(["Bonjour", ",", "ceci", "est", "une", "phrase",
"découpée", "en", "mots", "."], is_split_into_words=True).
```

Les modèles sont souvent pré-entraînés avec des tokeniseurs sous-lexicaux(*subword*), ce qui signifie que même si vos entrées ont déjà été divisées en mots, chacun de ces mots pourrait être à nouveau divisé par le tokeniseur. Prenons un exemple à cet effet :

Nous définissons d'abord un exemple, qui est la phrase à l'index 4 de l'ensemble d'entraînement. Ensuite, nous effectuons la tokenisation en divisant les mots en sous-mots. Enfin, nous imprimons les tokens obtenus comme ci-dessous.

```
['[CLS]', 'it', 'is', 'described', 'under', 'the', 'name',
'of', 'br', '##eit', '##horn', 'by', 'sa', '##uss', '##ure',
',', 'who', 'passed', 'a', 'couple', 'of', 'days', 'on',
'the', 'mont', 'ce', '##r', '##vin', 'and', 'ascended', 'an',
'inferior', 'horn', 'or', 'summit', 'which', 'he', 'calls',
'the', 'ci', '##me', 'br', '##une', 'du', 'br', '##eit',
'##horn', ',', 'round', 'which', 'we', 'made', 'a', 'circuit',
'and', 'which', 'we', 'left', 'far', 'below', 'us', 'in',
'our', 'ascent', '.', '[SEP]']
```

En effet, la tokenisation en sous-lexicaux est effectuée pour mieux traiter des textes de tailles différentes et pour gérer efficacement les mots inconnus (qui sont alors divisés en segments de mots faisant partie du vocabulaire d'entrainement). Cette approche est fréquemment utilisée en traitement du langage naturel, notamment lors de l'utilisation de modèles de langage pré-entraînés.

En général, les termes couramment employés demeurent inchangés, tandis que ceux moins fréquents peuvent être décomposés en plusieurs sous-mots. Cette démarche s'avère utile dans plusieurs cas : parfois, le modèle peut rencontrer des termes absents de son vocabulaire pré-entraîné, appelés mots hors vocabulaire. Grâce à la segmentation en sous-mots, ces termes peuvent être traités en utilisant les informations de leurs composants. De plus, cette

segmentation facilite la généralisation vers des mots et des expressions similaires, même si elles n'ont pas été observées durant l'entraînement. En outre, elle permet de traiter différentes formes et variantes d'un terme, comme les temps verbaux, les pluriels, les conjugaisons, etc. Cette approche contribue également à la réduction de la taille du vocabulaire du modèle, économisant ainsi des ressources de stockage et de calcul.

Dans l'exemple ci-dessus, les mots "*Breithorn*" et "*Saussure*" ont été divisés en trois sous-lexicaux en raison de leur rareté potentielle dans le modèle pré-entraîné. Cette situation nécessite un traitement spécifique de nos étiquettes, car les IDs d'entrée renvoyés par le tokeniseur sont plus longs que les listes d'étiquettes de notre ensemble de données. Cette extension découle des possibles subdivisions des mots en plusieurs tokens, en plus de l'ajout des tokens spéciaux, comme nous le constatons avec [CLS] et [SEP]. Ces deux étiquettes jouent des rôles spéciaux dans les modèles Transformer. L'étiquette [CLS] marque le début d'une phrase ou résume la sémantique de la phrase entière, principalement utilisée dans les tâches de classification pour encapsuler la représentation de la phrase complète. L'étiquette [SEP] sert à séparer différentes phrases ou segments d'entrée, permettant ainsi au modèle de les distinguer.

Pour résoudre le problème de l'incohérence entre les tokens en entrée et liste des étiquettes, nous devons aligner les tokens d'entrée prétraités avec leurs étiquettes correspondantes, garantissant ainsi que le modèle puisse calculer correctement les pertes pendant l'entraînement. Voici les étapes détaillées :

Nous utilisons d'abord la méthode `"tokenized_input.word_ids()"` pour récupérer la liste des `"word_id"` pour chaque token du texte d'entrée prétraité. Ces "word_id" représentent le mot auquel chaque token appartient, et les tokens spéciaux ont un "word_id" égal à `"None"`.

Ensuite, nous utilisons `"aligned_labels"` pour créer une liste d'étiquettes alignée comme illustré dans le pseudo-code ci-dessous :

```
1. Pour chaque i dans la liste des identifiants de mots de tokenized_input:
2.     Si i est None:
3.         Ajouter -100 à aligned_labels
4.     Sinon: ajouter example[f"{task}_tags"][i] à aligned_labels
5.     Fin Si
6. Fin Pour
```

Enfin, nous imprimons la longueur de `"aligned_labels"` ainsi que le nombre de `"tokenized_input["input_ids"]"`. Ces deux quantités devraient être égales, ce qui

garantit qu'il y a une étiquette correspondante pour chaque token d'entrée. Cela permettra au modèle de calculer correctement les pertes lors de l'entraînement.

Grâce à cette étape, nous nous assurons que les tokens d'entrée sont correctement alignés avec les étiquettes correspondantes, ce qui fournit au modèle des données d'entrée précises et des étiquettes cibles pour l'entraînement.

Il nous reste maintenant une dernière étape dans notre processus de prétraitement, qui consiste à intégrer les étapes que nous venons de décrire dans une seule fonction. Pour ce faire, nous définissons une fonction nommée "`tokenize_and_align_labels`" pour prétraiter nos exemples. Nous utilisons le tokeniseur avec les arguments "`truncation=True`" (pour tronquer les textes plus grands que la taille maximale autorisée par le modèle) et "`is_split_into_words=True`". Ensuite, nous alignons les étiquettes avec les IDs de tokens en utilisant la stratégie que nous avons choisie. En appliquant cette fonction à toutes les phrases de notre ensemble de données, nous utilisons simplement la méthode "`map`" de notre objet "`dataset`" créé précédemment. Cela prétraite toutes les données en une seule commande comme ci-dessous :

```
"tokenized_datasets=datasets.map(tokenize_and_align_labels
                , batched=True) "
```

Nous pouvons examiner les résultats d'une phrase spécifique de la manière suivante :

```
tokenized_datasets["train"]["labels"][0]
```

Cette commande nous permettra d'accéder aux étiquettes associées à la première phrase de l'ensemble d'entraînement (*train*). Le résultat est comme suit :

```
[-100, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 0, 1, 1, 1, 1, 1, 2, 2, 2,
                        2, 2, 2, -100]
```

Les résultats sont automatiquement mis en cache par la bibliothèque "`Datasets`" pour éviter de répéter cette étape la prochaine exécution du notebook. De plus, la bibliothèque Datasets détecte automatiquement si la fonction passée à "`map`" a changé et utilise donc les données mises en cache si possible.

En résumé, cette partie décrit comment prétraiter les données à l'aide du tokeniseur, notamment la tokenisation, l'alignement des étiquettes et la mise en cache des résultats, afin de préparer les données pour l'entraînement du modèle et l'analyse des données.

2.3. Effectuer des ajustements sur le modèle pré-entraîné

Dans cette section, notre objectif est d'affiner le modèle pré-entraîné pour une tâche spécifique de classification de tokens en utilisant les données prétraitées. En d'autres termes, notre objectif est de prédire si un token correspond à un oronyme ou s'il constitue un élément d'un oronyme. Si c'est le cas, nous cherchons également à déterminer de quel élément il s'agit précisément.

Pour commencer, nous constituons une liste d'étiquettes adaptée à la tâche et établissons une correspondance entre ces étiquettes et leurs indices respectifs. Ensuite, nous mettons en place un modèle de classification de tokens en sélectionnant des points de contrôle pré-entraînés appropriés et en ajustant les paramètres selon nos besoins. En parallèle, nous utilisons un optimiseur pour gérer l'optimisation des paramètres du modèle et nous compilons le modèle en vue de l'entraînement. Le passage des données prétraitées à un ensemble de données TensorFlow (Abadi et al., 2016)[41] permet d'alimenter le modèle en données d'entraînement. Pendant le processus d'apprentissage, nous définissons des fonctions de calcul des métriques pour évaluer la performance du modèle sur l'ensemble de validation. Cela nous permet de surveiller et d'améliorer en temps réel les performances du modèle à mesure qu'il apprend. Enfin, en utilisant l'ensemble de données d'apprentissage, nous affinons le modèle à travers des cycles itératifs d'apprentissage et en exploitant des fonctions de rappel. Tout cela vise à obtenir des résultats optimaux pour la tâche de classification de tokens spécifique. Nous expliquons ces étapes en détails dans la section suivante.

2.3.1. Construction du modèle

Maintenant que nos données ont été prétraitées, nous démarrons en construisant un modèle destiné à la classification des tokens à l'aide de la classe `"TFAutoModelForTokenClassification"` de la bibliothèque "Transformers". Tout d'abord, nous définissons une liste `"label_list"` contenant des étiquettes qui représentent différentes catégories de tokens, telles que le début d'un oronyme (B-LOC), l'intérieur d'un oronyme (I-LOC) et autre (O). Ensuite, nous créons deux dictionnaires `"id2label"` et

[41] TensorFlow : est une plateforme d'apprentissage automatique open-source développée par Google en 2015, spécialement conçue pour la création et l'entraînement de divers modèles d'apprentissage automatique, notamment ceux basés sur l'apprentissage profond. Cette plateforme offre une gamme étendue d'outils et de bibliothèques, permettant aux développeurs de concevoir des applications d'apprentissage automatique hautement performantes dans de nombreux domaines, allant de la reconnaissance d'images au traitement du langage naturel en passant par les systèmes de recommandation; Source : https://www.tensorflow.org/

"label2id" qui établissent la correspondance entre les étiquettes et leurs indices respectifs (0,1,2).

Par la suite, en appelant la méthode "from_pretrained" de la classe "TFAutoModelForTokenClassification", nous chargeons les pondérations appropriées du modèle pré-entraîné et ainsi créons un modèle destiné à la tâche de classification des tokens. Au cours de cette étape, le paramètre "num_labels" a été défini pour refléter le nombre de catégories d'étiquettes, garantissant que le modèle est informé du nombre d'étiquettes à prédire. De plus, les dictionnaires "id2label" et "label2id" ont été passés au modèle, assurant leur utilisation dans les étapes d'apprentissage et d'inférence ultérieures.

2.3.2. Construction de l'optimiseur

Après avoir exécuté l'étape précédente, nous pouvons observer un avertissement nous informant que nous supprimons certaines pondérations (les couches "vocab_transform" et "vocab_layer_norm") et en initialisons d'autres de manière aléatoire (les couches "pre_classifier" et "classifier"). Ceci est tout à fait normal dans ce cas, car nous retirons la tête utilisée pour pré-entraîner le modèle sur un objectif de modélisation de langage masqué, et nous la remplaçons par une nouvelle tête pour laquelle nous n'avons pas de pondérations pré-entraînées. Par conséquent, la bibliothèque nous avertit que nous devrions affiner ce modèle avant de l'utiliser pour l'inférence, ce que nous allons précisément faire.

Nous avons appelé précédemment la classe "TFAutoModelForTokenClassification", qui est conçue pour construire des modèles Keras (Chollet & others, 2015)[42] basés sur TensorFlow. Donc, dans cette étape nous utilisons l'interface Keras fournie par Transformers pour construire un modèle destiné à la classification de tokens. Ainsi, pour compiler un modèle Keras, nous devons définir un optimiseur et une fonction de perte. Pour ce faire, nous utilisons la fonction "create_optimizer".

Tout d'abord, nous définissons le nombre total d'époques d'entraînement ("num_train_epochs") à 3. Le nombre d'époques représente le nombre de fois que le modèle est mis à jour sur l'ensemble des données d'entraînement. Un nombre d'époques

[42] Keras : est une bibliothèque de réseaux neuronaux largement utilisée dans le domaine de l'apprentissage profond. Elle simplifie la création et l'entraînement de modèles de réseaux neuronaux, accélérant ainsi leur développement et leur évolutivité. Keras est compatible avec plusieurs frameworks d'apprentissage profond, ce qui en fait un outil puissant pour les tâches de construction de modèles complexes; Source : https://keras.io/

insuffisant peut empêcher le modèle d'apprendre complètement les caractéristiques des données, tandis qu'un nombre d'époques excessif peut conduire à un surapprentissage des données d'entraînement. En général, il est préférable de commencer avec un petit nombre d'époques pour un entraînement préliminaire, puis d'ajuster ce nombre en observant les performances du modèle sur un ensemble de validation. En plus, l'entraînement d'un modèle avec un grand nombre d'époques peut nécessiter beaucoup de temps et de ressources de calcul. Fixer un petit nombre d'époques peut permettre d'obtenir des résultats acceptables dans un temps limité.

Ensuite, nous calculons le nombre total d'étapes d'entraînement (`"num_train_steps"`) en divisant le nombre d'exemples dans l'ensemble de données d'entraînement par la taille de lot (*batch size*)[43], puis en multipliant par le nombre total d'époques. Dans notre cas, avec 7 855 exemples d'entraînement et une taille de lot (*batch size*) fixé à 16, chaque mise à jour des paramètres du modèle se fait avec 16 échantillons. Si une époque inclut l'ensemble des 7 855 exemples, alors chaque époque comprendra environ 490 mises à jour (7855 / 16 ≈ 490).

Et puis, nous utilisons la fonction `"create_optimizer"` pour créer un optimiseur et une décroissance du taux d'apprentissage.

L'optimiseur est un algorithme qui ajuste les poids du modèle en fonction de la perte (*loss function*) calculée pendant l'entraînement. Cela permet d'appliquer une stratégie d'optimisation personnalisée au modèle. La décroissance du taux d'apprentissage (*learning rate decay*) est une stratégie qui vise à réduire progressivement le taux d'apprentissage pendant l'entraînement, afin d'aider le modèle à converger plus stablement vers la solution optimale à mesure que l'entraînement progresse.

Le taux d'apprentissage est un hyperparamètre qui contrôle l'amplitude des mises à jour des paramètres du modèle. Un taux d'apprentissage élevé peut accélérer la convergence du modèle, mais peut également entraîner le saut par-dessus des *optima* locaux ; un taux d'apprentissage faible peut améliorer la stabilité, mais peut ralentir la vitesse d'entraînement.

Nous définissons le taux d'apprentissage initial (`"init_lr"`) à 2e-5 (0,00002), il s'agit de la vitesse à laquelle le modèle ajuste ses paramètres en fonction de la perte calculée lors de la mise à jour. Un taux d'apprentissage initial typique pour les modèles de traitement du langage

[43] Batch size : la taille du lot indique le nombre d'échantillons traités lors d'une mise à jour du modèle. Une taille de lot plus grande peut tirer parti de l'accélération matérielle et réduire la variance des mises à jour de paramètres, ce qui peut parfois accélérer l'entraînement. Cependant, une taille de lot trop grande peut causer des problèmes de mémoire et de ressources informatiques.

naturel est généralement faible, comme 2e-5. Cela permet un apprentissage progressif et une convergence plus stable. En ce qui concerne le nombre total d'étapes d'entraînement (`"num_train_steps"`), nous l'avons déjà calculé précédemment. Pour le taux de décroissance des poids (`"weight_decay_rate"`), il est fixé à 0,01. Ce dernier indique que les poids élevés sont réduits de 1 % à chaque mise à jour. Cette valeur introduit une pénalité pour les poids élevés dans le modèle, contribuant ainsi à sa régularisation et à la prévention du surapprentissage. Quant au nombre d'étapes de préchauffage (`"num_warmup_steps"`), il est fixé à 0. Cela signifie que cette phase n'est pas utilisée. Ce paramètre se rapporte aux premières étapes d'entraînement, au cours desquelles le taux d'apprentissage augmente progressivement pour aider le modèle à converger plus rapidement.

Ces paramètres sont généralement utilisés comme point de départ pour l'entraînement de modèles d'apprentissage profond dans des tâches de traitement du langage naturel en raison de leur efficacité prouvée.

Le but de cette étape est de préparer un optimiseur pour l'entraînement du modèle et de configurer une décroissance du taux d'apprentissage. Cela permettra d'ajuster progressivement le taux d'apprentissage pendant l'entraînement du modèle, contribuant ainsi à optimiser les performances de l'apprentissage.

Enfin, nous compilons le modèle que nous avons créé précédemment en utilisant la méthode `"compile"`. Cette étape de compilation prépare le modèle pour l'entraînement en définissant l'optimiseur à utiliser pour ajuster les poids du modèle et minimiser la perte lors de la phase d'apprentissage.

2.3.3. Construction d'un collecteur de données et ensembles de données

Nous avons maintenant besoin d'un collecteur de données pour fusionner par lots nos exemples traités tout en appliquant un remplissage (*padding*) pour leur donner la même taille (on rajoute des tokens vides afin de calibrer tous les exemples à la longueur de l'exemple le plus long). À cette fin, nous utilisons `"DataCollatorForTokenClassification"` pour cette tâche car il peut alimenter non seulement les données d'entrée, mais aussi les étiquettes. De plus, notre collecteur de données a été conçu pour être compatible avec divers *frameworks*, nous devons donc nous assurer que le paramètre `"return_tensors='np'"` est défini pour récupérer les tableaux NumPy (Harris et al., 2020)[44]. En effet, dans notre pipeline `"TF dataset"`,

[44] NumPy : est une bibliothèque de programmation open source de tableaux pour le langage Python. Elle joue un rôle essentiel dans les pipelines d'analyse de la recherche dans des domaines variés. Dans le domaine de

nous utilisons le chargeur NumPy en interne et nous l'enveloppons avec "tf.data.Dataset" à la fin. Par conséquent, l'utilisation du paramètre "np" est généralement plus fiable et performant dans ce contexte.

Ensuite, nous convertissons nos ensembles de données en "tf.data.Dataset", un format que Keras peut comprendre naturellement. Il existe deux façons de le faire : nous pouvons utiliser la méthode légèrement plus bas niveau "Dataset.to_tf_dataset()", ou bien nous pouvons utiliser "Model.prepare_tf_dataset()". La différence fondamentale entre ces deux méthodes réside dans le fait que la méthode Model peut analyser le modèle afin de déterminer automatiquement quels noms de colonnes peuvent être utilisés en tant qu'entrée. Cela signifie qu'il n'est pas nécessaire de les spécifier manuellement.

Nous créons ensuite deux ensembles de données, l'un pour l'entraînement et l'autre pour la validation. Pour cela, nous utilisons la méthode "model.prepare_tf_dataset()". Pour le premier ensemble (train_set), nous utilisons les données prétraitées de l'ensemble d'entraînement (tokenized_datasets["train"]). Nous mélangeons les exemples (shuffle=True), définissons la taille des lots (batch_size) et utilisons le collecteur de données (data_collator) que nous avons configuré précédemment pour organiser les exemples en lots en appliquant le remplissage.

De même, pour le deuxième ensemble (validation_set), nous utilisons les données prétraitées de l'ensemble de validation (tokenized_datasets["validation"]). Cette fois-ci, nous ne mélangeons pas les exemples (shuffle=False), mais nous utilisons à nouveau la même taille de lot et le même "data collator" pour la création des lots.

En somme, cette étape de la préparation des ensembles de données est cruciale pour l'entraînement et la validation du modèle, car elle organise les données dans un format compréhensible par Keras et les prépare pour le processus d'apprentissage.

2.3.4. Métriques d'évaluation

À ce stade, il est temps d'aborder la question des métriques. Le *framework* Seqeval nous fournit un ensemble intéressant de métriques telles que l'exactitude (*accuracy*), la précision, le

l'apprentissage automatique et de l'analyse de données, NumPy est largement utilisé pour traiter et manipuler des données, effectuer une variété d'opérations mathématiques et créer et gérer des structures de données multidimensionnelles; Source : https://numpy.org/

rappel et le score F1. Ces quatre métriques sont des indicateurs couramment utilisés pour évaluer les performances des modèles de classification.

Parmi eux, l'exactitude mesure le rapport entre le nombre d'échantillons correctement prédits par le modèle et le nombre total d'échantillons. C'est l'indicateur le plus intuitif, mais dans le cas de jeux de données déséquilibrés, une exactitude élevée peut masquer les problèmes de performances du modèle sur les classes minoritaires.

La précision représente le rapport entre les échantillons prédits positifs par le modèle et le nombre réel d'échantillons positifs. Elle se concentre sur la proportion de vrais positifs parmi les prédictions positives du modèle, et convient aux cas où les faux positifs sont moins tolérés.

Le rappel indique le rapport entre tous les échantillons positifs réels et les échantillons correctement prédits positifs par le modèle. Il met en évidence la capacité du modèle à capturer les échantillons positifs réels, et est adapté aux situations où les faux négatifs sont moins acceptables.

Le score F1 est une métrique combinée de la précision et du rappel, calculée comme la moyenne harmonique de ces deux valeurs. Il permet d'équilibrer les performances du modèle en termes de précision et de rappel, et est particulièrement utile dans les cas de jeux de données déséquilibrés.

La prise en compte de ces métriques permet d'évaluer les performances du modèle et de choisir les stratégies d'entraînement ainsi que les configurations d'hyperparamètres appropriées en fonction du problème spécifique.

Maintenant, tout ce que nous avons à faire est de transmettre certaines prédictions et étiquettes à Seqeval. Pour ce faire, nous utilisons la fonction *callback* `"KerasMetricCallback"` que nous calculons sur les valeurs de validation du jeu de données à chaque époque, puis nous l'affichons et enregistrons les valeurs retournées, ce qui sera utilisé par d'autres *callback* tels que `"TensorBoard`[45]`"` et `"EarlyStopping`[46]`"`, ce qui permet une plus grande flexibilité avec les fonctions de calcul des métriques.

[45] TensorBoard : est un outil de visualisation et de suivi du processus d'apprentissage profond, généralement utilisé en conjonction avec TensorFlow. Il peut montrer les tendances des données d'entraînement et de validation, permettant aux utilisateurs de détecter le surapprentissage ou le sous-apprentissage et d'ajuster les hyperparamètres pour optimiser les performances du modèle.

[46] EarlyStopping : est une technique de régularisation visant à prévenir le surapprentissage. Elle surveille les performances du modèle sur l'ensemble de validation pendant l'entraînement. Lorsque les performances sur l'ensemble de validation cessent de s'améliorer ou commencent à se dégrader, EarlyStopping arrête l'entraînement, évitant ainsi le surapprentissage du modèle aux données d'entraînement.

À cet effet, nous chargeons d'abord la métrique "seqeval" de la bibliothèque "datasets" en utilisant la fonction "load_metric". Cette métrique est utilisée pour calculer des mesures de précision, de rappel, de F1 et d'exactitude spécifiques à la classification de séquences.

Ensuite, nous utilisons la métrique chargée pour calculer les métriques en utilisant les prédictions et les références d'exemples de validation. Cela nous donne une mesure de base de ces métriques pour une première compréhension des performances du modèle.

Après cela, nous définissons la fonction "compute_metrics". Cette fonction prend les prédictions et les étiquettes en entrée et effectue les étapes nécessaires pour calculer les métriques d'intérêt. Les prédictions sont transformées en indices des classes les plus probables en utilisant "np.argmax". Pendant cette démarche, nous excluons toutes les prédictions associées à l'étiquette -100, ce qui signifie une absence d'étiquette. Cela peut se produire lorsque le token n'a pas d'étiquette attribuée, ou lorsqu'il s'agit d'un token de remplissage. Cette exclusion vise à éviter d'intégrer les tokens de remplissage ou non étiquetés dans le calcul des mesures de performance.

Par la suite, en employant les prédictions transformées et les étiquettes nettoyées, nous recourons à la métrique "seqeval" pour calculer les valeurs finales des métriques d'exactitude, de précision, de rappel, et de F1.

Enfin, nous utilisons le "KerasMetricCallback" pour créer un *callback* de métrique. Ce rappel utilisera la fonction "compute_metrics" que nous avons définie pour calculer les métriques à chaque époque en utilisant l'ensemble de validation (validation_set).

En résumé, cette fonction illustre la création et l'utilisation d'un rappel de métrique personnalisé pour calculer et surveiller les performances du modèle sur l'ensemble de validation pendant l'entraînement.

2.3.5. Mise en œuvre de l'entraînement

Nous atteignons à présent notre dernière étape. Dans cette phase, nous allons commencer à entraîner notre modèle et faire quelques réglages supplémentaires pour un entraînement optimal.

Tout d'abord, nous importons le "PushToHubCallback" depuis "transformers.keras_callbacks" et le "TensorBoard" depuis

"tensorflow.keras.callbacks". Ensuite, nous extrayons le nom du modèle à partir du chemin "model_checkpoint".

Par la suite, nous définissons un identifiant unique pour le modèle dans le Hub, en utilisant le nom du modèle et en ajoutant "-finetuned-{task}". Cet identifiant sera employé pour téléverser le modèle dans le Hub des modèles de Hugging Face.

Nous instaurons un *callback* "TensorBoard" pour enregistrer les journaux d'entraînement dans le dossier "./tc_model_save/logs". "TensorBoard" constitue un outil puissant de visualisation du processus d'apprentissage, qui nous aide à suivre la progression de l'apprentissage et les performances de notre modèle comme illustrée dans la figure 9 :

```
Epoch 1/3
  6/490 [..............................] - ETA: 1:53 - loss: 0.9633WARNING:tensorflow:Callback method `on_train_batch_end` is slow compared to the batch time (bat
490/490 [==============================] - 163s 309ms/step - loss: 0.1248 - val_loss: 0.0642 - precision: 0.6858 - recall: 0.7660 - f1: 0.7237 - accuracy: 0.9752
Epoch 2/3
490/490 [==============================] - 154s 314ms/step - loss: 0.0513 - val_loss: 0.0510 - precision: 0.7684 - recall: 0.8323 - f1: 0.7991 - accuracy: 0.9812
Epoch 3/3
490/490 [==============================] - 153s 311ms/step - loss: 0.0326 - val_loss: 0.0499 - precision: 0.7896 - recall: 0.8506 - f1: 0.8189 - accuracy: 0.9826
<keras.callbacks.History at 0x7f55b4397c10>
```

Figure 8. Visualisation du processus d'apprentissage

Ensuite, nous recourons au "PushToHubCallback" pour créer un callback chargé de téléverser le modèle vers le Hub. Nous lui fournissons un répertoire de sortie, un tokeniseur et un identifiant pour le modèle du Hub. Ainsi, une fois l'entraînement achevé, notre modèle sera automatiquement téléchargé dans le Hub.

Enfin, nous regroupons tous les *callbacks* en une liste qui est passée comme argument à la fonction "fit()". Dans cette fonction, nous transmettons l'ensemble d'entraînement, l'ensemble de validation, le nombre d'époques d'entraînement et la liste de *callbacks*. La fonction "fit()" lancera le processus d'entraînement du modèle et déclenchera les *callbacks* à la fin de chaque époque pour enregistrer le processus d'entraînement et téléverser le modèle.

Pour résumer, cette étape marque la conclusion de l'intégralité du processus, couvrant l'entraînement du modèle, l'enregistrement et le téléchargement du modèle vers le Hub des modèles pour assurer une formation et un partage efficaces de notre modèle.

2.4. Évaluation du modèle DistilBERT

Pour évaluer la performance réelle du modèle, nous utilisons ce dernier pour effectuer la reconnaissance des oronymes sur les phrases du fichier "list_without_oronymes.txt" que nous avons obtenu dans la section "Prétraitement

du corpus" précédente. Les entités géographiques reconnues seront ensuite écrites dans un fichier nommé "`results.txt`". Les étapes spécifiques sont les suivantes :

Nous commençons par l'importation des bibliothèques nécessaires, notamment la bibliothèque Transformers (pour charger le modèle et le tokeniseur), TensorFlow (pour le traitement de données), NumPy (pour les opérations numériques) et jsonlines (pour le traitement de fichiers texte). Ensuite, un modèle et un tokeniseur adaptés à la tâche de classification de tokens sont chargés à partir du modèle pré-entraîné "`distilbert-base-uncased-finetuned-ner`".

Ensuite, nous utilisons la variable "`corpus_file_path`" pour lire le fichier texte nommé "`list_without_oronymes.txt`". Ce fichier contient des phrases que le modèle n'a pas rencontrées pendant l'entraînement. Ces phrases serviront à l'inférence et à la reconnaissance des entités. Pour chaque phrase, elle sera d'abord tokenisée en utilisant le tokeniseur. Et puis, il effectue des prédictions sur les token de la phrase en utilisant la méthode call du modèle (équivalente à "`model(tokenized).logits`"). Les sorties du modèle sont des `logits`, qui représentent les scores associés à chaque classe pour chaque token.

Ainsi, pour chaque token, la classe prédite est déterminée en choisissant celle avec le score le plus élevé en utilisant la fonction "`np.argmax`". Ensuite, les indices de classe correspondants sont récupérés.

Par la suite, les tokens tokenisés et les étiquettes prédites sont combinés sous forme de paires (token, étiquette). Ces paires représentent les prédictions de classification et les étiquettes associées pour chaque token. Pour chaque paire token-étiquette prédite, nous vérifions si l'étiquette correspond à une entité géographique ("B-LOC" ou "I-LOC"). Si c'est le cas, le token est ajouté à une liste "`filtered_tokens`".

Finalement, pour chaque phrase, nous prenons les tokens d'oronymes reconnus, les joignons avec des espaces, et les écrivons dans le fichier "`results.txt`". Ce fichier contient tous les oronymes identifiés dans le "`list_without_oronymes`" pour des analyses ultérieures.

3. *Modèle de référence - spaCy*[47]

Afin de comparer les performances de notre modèle avec celles d'autres modèles, nous avons également effectué des tests avec le fichier "sous-corpus" à l'aide de spaCy (Montani et al., 2023). spaCy est une bibliothèque open-source de traitement du langage naturel (TAL) largement utilisée, développée en 2015 par Matthew Honnibal et Ines Montani. Elle propose un ensemble de fonctionnalités puissantes, notamment la désambiguïsation, l'annotation lexicale, l'analyse syntaxique, la reconnaissance d'entités nommées, et bien plus encore. spaCy est reconnue pour sa grande efficacité et convient parfaitement au traitement à grande échelle de données textuelles en langage naturel.

Nous commençons par installer les dépendances nécessaires en exécutant les commandes suivantes pour mettre à jour pip, setuptools et wheel :

```
! pip install -U pip setuptools wheel
```

Ensuite, installons la bibliothèque spaCy en exécutant la commande :

```
! pip install -U spacy
```

Après avoir installé spaCy, nous procédons au téléchargement du modèle de base pour l'anglais, `"en_core_web_sm"`. Ce modèle contient des fonctionnalités de traitement linguistique spécifiques pour le texte en anglais.

Ensuite, nous configurons spaCy pour gérer de longs textes en augmentant la limite de traitement à 9 000 000 de caractères. Cela permet de s'assurer que le traitement peut gérer des textes plus longs. Et puis nous spécifions le chemin du fichier texte à traiter `"corpus_file_path"` ainsi que la taille de chaque bloc de traitement `"block_size"`.

Par la suite, nous utilisons une boucle pour lire le fichier texte et traiter son contenu par blocs. La taille de chaque bloc est spécifiée par la variable `"block_size"`. Nous fusionnons le contenu de chaque bloc en une seule grande chaîne de caractères `"corpus"`. Ensuite, nous passons cette grande chaîne de caractères au pipeline de traitement `"nlp"` de spaCy, qui effectuera la tokenisation, la reconnaissance d'entités, etc. Pour sauvegarder les résultats, nous définissons le chemin du fichier de sortie `"output_file_path"`, où les résultats de la reconnaissance d'entités seront enregistrés.

[47] Source : https://spacy.io/

Finalement, nous définissons une fonction "process_blocks", qui prend en entrée le chemin du fichier texte, la taille des blocs et le chemin du fichier de sortie. À l'intérieur de la fonction, nous ouvrons le fichier d'entrée et lisons son contenu par blocs dans une boucle. Pour chaque bloc, nous utilisons spaCy pour effectuer la reconnaissance d'entités et écrivons les entités et leurs étiquettes correspondantes dans le fichier de sortie.

Nous pouvons également utiliser le module "displacy" de spaCy pour visualiser les résultats de la reconnaissance d'entités dans un format annoté.

4. Modèle de référence - BERT-base (ci-après BERT)

4.1. Entraînement du modèle

Étant donné que DistilBERT est une version réduite du modèle BERT, bien que ses performances n'aient pas été sensiblement compromises, nous souhaitons tout de même évaluer comment le modèle BERT se comportent sur nos données. Le processus d'entraînement demeure identique à celui du modèle DistilBERT précédent, et nous l'avons de nouveau implémenté sur Colab. La seule variation réside dans l'ajustement des données d'entraînement, et les étapes spécifiques sont détaillées ci-dessous.

Pour DistilBERT, notre ensemble d'entraînement était composé de seulement 11 222 phrases contenant des occurrences liées aux oronymes énumérés dans la liste d'origine. En revanche, pour BERT-base, nous avons enrichi cet ensemble d'entraînement en y ajoutant 1000 phrases dépourvues d'occurrences d'oronymes. Ces nouvelles phrases ont été sélectionnées manuellement à partir des prédictions générées par DistilBERT dans le fichier "results.txt" (tous ces résultats sont détaillés dans le chapitre suivant). Ce fichier est issu du fichier "list_without_oronymes.txt", comme nous utiliserons ce sous corpus pour évaluer le modèle BERT-base ultérieurement, nous avons également retiré les 1000 phrases choisies du fichier "list_without_oronymes.txt".

Ensuite, nous avons annoté ces 1000 nouvelles phrases avec l'étiquette "2" (O dans le format BIO) pour chaque token, car elles ne contiennent pas d'oronyme. Pour réaliser cela, nous avons utilisé une fonction similaire à celle que nous avons utilisée précédemment. Nous commençons par définir une fonction appelée "process_sentences". Cette fonction prend en entrée un fichier contenant des phrases non traitées et enregistre les phrases traitées au format JSONL dans un fichier de sortie. Le processus de traitement est le suivant :

Nous lisons les phrases du fichier d'entrée ligne par ligne.

Chaque phrase est divisée en mots (tokens) en utilisant la fonction "word_tokenize" de la bibliothèque NLTK (Natural Language Toolkit).

Pour chaque phrase traitée, toutes les balises sont attribuées avec la valeur "2".

Les informations concernant le traitement de chaque phrase sont ensuite écrites dans le fichier de sortie au format JSONL en utilisant la bibliothèque "jsonlines". Ces informations incluent des identifiants uniques générés en fonction de l'index de la phrase, ainsi que les tokens et les balises NER correspondantes.

Cette série de 1000 phrases commence avec l'identifiant "nominalp#11223" pour s'assurer que nous pourrons ensuite fusionner ces données dans l'ensemble d'entraînement utilisé pour DistilBERT, puis les répartir de nouveau.

Pour segmenter ces nouvelles données d'entraînement, nous avons appliqué la même fonction que celle décrite dans la section 1.4 de ce chapitre. De cette manière, nous avons obtenu un total de trois sous-ensembles distincts :

"bert_train.jsonl" comprenant 8555 phrases, soit 70% du total,

"bert_validation.jsonl" comprenant 2444 phrases, équivalant à 20%,

"bert_test.jsonl" contenant 1223 phrases, soit 10%.

Plus précisément, dans le sous-ensemble "train", 717 des 1000 phrases nouvellement ajoutées sont incluses, le sous-ensemble "validation" contient 182 phrases, et le sous-ensemble "test" contient 101 phrases.

Les étapes précédentes représentent le processus de préparation des nouvelles données d'entraînement. Pour l'entraînement du nouveau modèle BERT-base, nous devons simplement spécifier le paramètre "model_checkpoint" en tant que "bert-base-uncased". Toutes les autres étapes suivent le même processus que l'entraînement de DistilBERT.

4.2. Évaluation du modèle

Lors de l'évaluation du modèle BERT que nous venons de former, nous suivons la même méthodologie que celle employée pour évaluer le modèle DistilBERT. Cependant, la seule différence réside dans l'utilisation d'un nouveau fichier "new_subcorpus.txt". Ce fichier exclut délibérément les 1000 phrases que nous avions ajoutées aux données d'entraînement lors de la phase d'entraînement du modèle. Cette démarche vise à assurer l'objectivité et la précision

des résultats d'évaluation en veillant à ce que le modèle soit confronté à des données auxquelles il n'a pas été exposé durant son apprentissage.

Chapitre 3. Résultats et analyses

Dans ce dernier chapitre, nous allons présenter les résultats obtenus après avoir soumis notre corpus aux trois modèles distincts que nous avons introduits précédemment. Nous allons également examiner les métriques clés telles que l'exactitude (*accuracy*), la précision, le rappel et le score F1 pour chaque modèle. En plus de présenter les résultats, nous allons procéder à des analyses pour mieux comprendre les performances et les comportements de chaque modèle dans le contexte de la tâche de reconnaissance d'oronymes dans un grand corpus. Cette évaluation nous permettra de tirer des conclusions éclairées sur les avantages et les limites de chaque modèle, ainsi que sur les enseignements tirés de ces résultats pour des applications futures.

1. Résultats

Nous expliquerons successivement dans les sections suivantes les résultats et les performances des trois modèles DistilBERT, spaCy, et BERT.

1.1. Résultats du modèle DistilBERT

Pendant le processus d'entraînement, nous avons mis en place une fonction appelée `"compute_metrics"`. Cette fonction nous a permis d'obtenir les scores du modèle à la fin de chaque cycle d'entraînement en fonction des différentes métriques calculées sur l'ensemble de validation. En parallèle, nous avons également procédé à des évaluations en utilisant l'ensemble de test, et les résultats sont présentés de manière détaillée dans les Tableaux 3 et 4.

Train Loss	Validation Loss	Train Precision	Train Recall	Train F1	Train Accuracy	Epoch
0.1248	0.0642	0.6858	0.7660	0.7237	0.9752	1
0.0513	0.0510	0.7684	0.8323	0.7991	0.9812	2
0.0326	0.0499	0.7896	0.8506	0.8189	0.9826	3

Tableau 3. Métriques d'évaluation et historique des pertes de DistilBERT basées sur l'ensemble de validation

Train Loss	Validation Loss	Train Precision	Train Recall	Train F1	Train Accuracy	Epoch
0.1220	0.0725	0.6701	0.7073	0.6882	0.9707	1
0.0509	0.0585	0.7617	0.7839	0.7726	0.9777	2
0.0322	0.0584	0.7851	0.8266	0.8053	0.9792	3

Tableau 4. Métriques d'évaluation et historique des pertes de DistilBERT basées sur l'ensemble de test

Par ailleurs, nous avons appliqué le modèle pour effectuer des prédictions sur le fichier "`list_without_oronymes.txt`", composé de 63 355 phrases qui n'ont jamais été exposées au modèle pendant la phase d'entraînement et qui n'ont pas été annotées. Cela nous fournit une compréhension plus solide de ses performances en situation réelle. En plus, notre objectif de l'entraînement du modèle était de le doter de la capacité à identifier et à étiqueter les oronymes au sein du texte. Si le modèle est capable d'identifier les oronymes dans le nouveau corpus sans avoir été exposé à ces données spécifiques pendant l'entraînement, cela confirme son aptitude à traiter des cas réels et inconnus, et cela valide sa pertinence pour des tâches pratiques.

Initialement, nous avons généré un total de 63 308 résultats à partir du processus de prédiction. Cependant, au cours de l'analyse, nous avons identifié la présence de nombreux bruits, avec le token [SEP] se démarquant par son apparition fréquente, comptabilisant 62 718 occurrences. Ce token [SEP] apparaissait à la fois de manière isolée et associée à d'autres tokens, comme nous pouvons constater dans la figure 9 :

`the valley of tri##ent [SEP]`

À part cela, il inclut une occurrence avec [CLS], qui était l'unique instance de ce type.

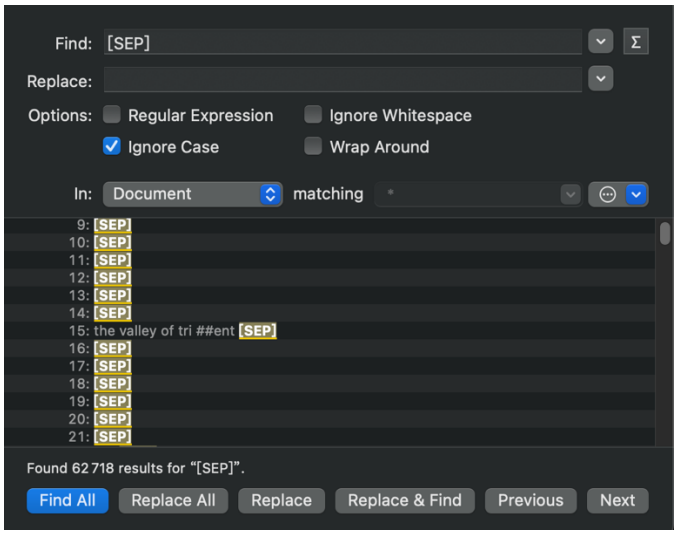


Figure 9. Nombre de token [SEP] dans les résultats du DistilBERT

Au-delà de ces balises spéciales, d'autres erreurs ont également été repérées. Afin d'obtenir une vue plus claire des résultats corrects, nous avons effectué un nettoyage manuel. Le processus se déroule comme suit :

La première étape consiste à utiliser une fonction "`process_line`" en Python pour supprimer tous les caractères spéciaux tels que [SEP], [CLS] ainsi que les caractères ## présents dans certains résultats à la suite de la tokenisation en sous-lexicaux par le tokeniser, par exemple,

"the valley of tri##ent [SEP]" devient "the valley of trient". nous éliminons les lignes vides supplémentaires résultant des deux premières opérations.

Dans la deuxième étape, nous utilisons la fonction "rechercher et remplacer" de l'éditeur de texte pour nettoyer les résultats contenant des apostrophes, lesquels sont entourés d'espaces avant et après l'apostrophe. Par exemple, "the val d ' entre mont". De la même manière, certains résultats contenant des points présentent ainsi un espace avant et après ces derniers, par exemple, "the val de st . marcel", nous devons donc supprimer ces espaces.

Dans la troisième étape, après avoir effectué les opérations précédentes, nous utilisons un programme Python pour compter les occurrences des résultats, en veillant à ne conserver qu'une seule occurrence des doublons. Pour ce faire, nous utilisons un compteur de Python (Counter) qui nous permet de suivre le nombre d'apparitions de chaque instance. Nous ouvrons le fichier contenant les résultats et parcourons chaque ligne. Nous supprimons ensuite les lignes en double, ne conservant qu'une seule occurrence de chaque ligne. Dernièrement, nous écrivons le contenu unique et leur nombre d'occurrence total dans le fichier de sortie. Après cela, il nous reste 9 468 résultats.

À partir de l'étape 4, nous entamons une phase de travail manuel essentielle. Dans cette phase, nous entreprenons le filtrage des résultats pour conserver en premier temps ceux qui sont les plus susceptibles d'être des oronymes, tels que *mont*, *vallée*, *val*, *glacier*, *col*, *-joch*, *-thal*, monte, *pic*, *pass*, *passo*, *aiguille*, *point*, *-spitz*, *-berg* et d'autres. Pour les cas incertains, comme quand il s'agit de noms propres, nous avons mené des recherches en ligne en utilisant les moteurs de recherche comme Google pour prendre des décisions éclairées sur leur inclusion ou exclusion. En revanche, si les éléments identifiés sont des rivières, des lacs ou des villages, personnes, nous les excluons, par exemple, de nombreux résultats en allemand se terminant par berg ne correspondent pas à un oronyme mais à un village, comme *Kranzberg*.

Enfin, nous avons supprimé les tokens qui n'ont pas de signification. Il s'agit généralement de résultats qui ne contiennent que quelques lettres, par exemple, "mo##e", "##aw", "##e", "ty", "to son", et de résultats incomplets comme "the", "glacier" etc. Par ailleurs, les tokens qui n'ont pas de lien avec les oronymes, par exemple, *family*, *house of*, *pine*, et *norway* etc., ont été également supprimés.

Suite à ce processus de nettoyage, nous avons réussi à réduire les résultats à une liste plus concise et significative, comprenant 3 346 entrées valides.

1.2. Résultats de spaCy

Avec spaCy, nous avons initialement obtenu un total de 83 237 résultats. Cependant, ce résultat contient un nombre important de prédictions incorrectes. Normalement, nous nous intéressons uniquement aux résultats étiquetés comme étant des lieux (LOC). Cependant, après plusieurs essais, nous avons constaté que spaCy attribue de manière incorrecte les véritables oronymes à d'autres catégories d'entités, et que de nombreuses entités étiquetées en tant que LOC ne correspondent pas non plus à des oronymes. Par conséquent, nous avons choisi de conserver l'ensemble des résultats identifiés par spaCy, ce qui explique pourquoi ce résultat est nettement plus élevé que celui obtenu avec DistilBERT avant le nettoyage.

Une fois de plus, nous avons entamé le processus de suppression des éléments en double. À l'issue de cette étape, nous avons obtenu un total de 28 695 résultats. Parmi ces résultats, nous avons identifié 18 types d'étiquette d'entités. Après avoir éliminé les doublons, et les erreurs, tout comme ce que nous avons fait pour les résultats du DistilBERT, nous avons conservé 3 666 résultats. Les données spécifiques pour chaque étiquette sont présentées dans le tableau ci-dessous :

Type d'entités	Quantité initiale	Qté après élimination des doublons	Qté après élimination des erreurs
ORG	12 235	5483	1320
PERSON	16 417	6979	852
LOC	1987	1048	561
FAC	1339	937	475
GPE	7641	2517	186
PRODUCT	1082	580	132
WORK_OF_ART	297	234	69
EVENT	132	95	37
LAW	49	43	11
Non classifié	3551	2613	23
CARDINAL	14 911	1936	0
DATE	8150	2490	0
TIME	6179	1719	0
NORP	3395	748	0
ORDINAL	3278	67	0
QUANTITY	2026	1017	0
LANGUAGE	295	11	0
MONEY	233	155	0
PERCENT	40	23	0
Nombre total	83 237	28 695	3666

Tableau 5. Détails des résultats de spaCy après les différentes étapes du traitement

1.3. Résultats du modèle BERT

La même fonction "`compute_metrics`" a été utilisée pour obtenir ces résultats, de manière similaire à celle utilisée pour le modèle DistilBERT. Les résultats basés sur l'ensemble de validation et l'ensemble de test sont illustrés dans les deux tableaux suivants :

Train Loss	Validation Loss	Train Precision	Train Recall	Train F1	Train Accuracy	Epoch
0.1034	0.0641	0.6823	0.8230	0.7461	0.9751	1
0.0419	0.0433	0.8160	0.8499	0.8326	0.9836	2
0.0229	0.0465	0.8265	0.8702	0.8478	0.9850	3

Tableau 6. Métriques d'évaluation et l'historique des pertes du BERT basées sur l'ensemble de validation

Train Loss	Validation Loss	Train Precision	Train Recall	Train F1	Train Accuracy	Epoch
0.1028	0.0577	0.7179	0.8009	0.7571	0.9775	1
0.0401	0.0461	0.7985	0.8540	0.8253	0.9833	2
0.0219	0.0468	0.8225	0.8738	0.8474	0.9852	3

Tableau 7. Métriques d'évaluation et l'historique des pertes du BERT basées sur l'ensemble de test

Pour évaluer le modèle BERT sur un corpus auquel il n'a pas été exposé pendant son apprentissage, contrairement au processus de validation de DistilBERT, nous utilisons une nouvelle version de `"list_without_oronymes"`, à savoir `"new_list_without_oronymes.txt"`, qui ne contient pas les 1000 phrases que nous avons utilisées pour l'entraînement du modèle. Cela nous a donné un total de 23 652 résultats. Parmi ces résultats, nous avons observé la présence de 13 655 tokens spéciaux [SEP], dont 2 577 étaient accompagnés d'autres tokens. De plus, nous avons repéré 13 tokens spéciaux [CLS]. Afin d'obtenir une liste de résultats plus nette, nous avons éliminé ces balises spéciales, aboutissant ainsi à 12 564 résultats. Ensuite, nous avons supprimé les doublons, ce qui nous a permis de réduire le total à 7 444 résultats.

Finalement, en appliquant les mêmes étapes de nettoyage que celles effectuées pour les résultats de DistilBERT, nous obtenons un total de 3 408 résultats à l'issue de cette procédure.

1.4. Intersection des résultats des trois modèles

Nous employons une fonction `"read_file_to_set"` pour comparer les mêmes parties des résultats des trois modèles DistilBERT, BERT et spaCy, les détails de la fonction sont montrés dans le pseudo-code ci-dessous :

```
 1. fonction lire_fichier_vers_ensemble(nom_du_fichier):
 2.     Ouvrir nom_du_fichier en mode lecture
 3.         Lire chaque ligne dans un ensemble
 4.         (Note: Si le nom_du_fichier est c, alors convertir chaque ligne en
minuscules)
 5.     Renvoyer cet ensemble
```

```
 6.
 7. Définir une fonction sauver_ensemble_vers_fichier(nom_du_fichier,
ensemble_de_données):
 8.     Ouvrir nom_du_fichier en mode écriture
 9.         Pour chaque élément dans ensemble_de_données:
10.             Écrire cet élément dans le fichier et passer à la ligne suivante
11.
12. Lire le fichier a dans l'ensemble a_set
13. Lire le fichier b dans l'ensemble b_set
14. Lire le fichier c dans l'ensemble c_set
15.
16. Calculer l'intersection de a_set et b_set et nommer le résultat ab_common
17. Calculer l'intersection de a_set et c_set et nommer le résultat ac_common
18. Calculer l'intersection de b_set et c_set et nommer le résultat bc_common
19. Calculer l'intersection de a_set, b_set et c_set et nommer le résultat
abc_common
20.
21. Sauver ab_common dans le fichier 'ab_common.txt'
22. Sauver ac_common dans le fichier 'ac_common.txt'
23. Sauver bc_common dans le fichier 'bc_common.txt'
24. Sauver abc_common dans le fichier 'abc_common.txt'
25.
```

Le fichier a correspond aux résultats nettoyés de DistilBERT, le fichier b aux résultats nettoyés de BERT et le fichier c aux résultats nettoyés de spaCy. Comme les résultats de spaCy contiennent des lettres majuscules, nous les avons tous convertis en lettres minuscules au cours du traitement afin de les comparer avec a et b. Les résultats détaillés sont présentés dans la figure ci-dessous :

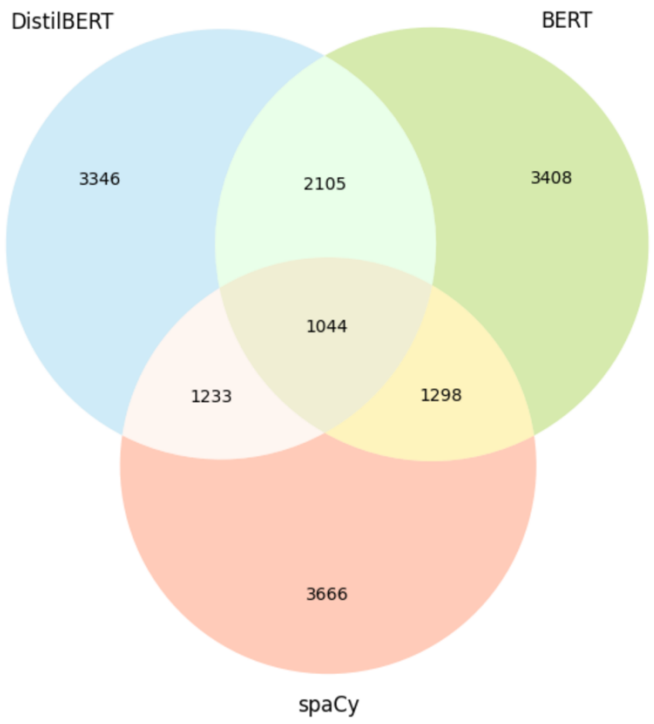


Figure 10. Intersection des données entre les résultats des 3 modèles

Nous pouvons voir sur la figure que DistilBERT et BERT ont entre eux plus de résultats identiques (2105)[48], DistilBERT et spaCy (1233)[49], BERT et spaCy (1298)[50], et les trois modèles ensemble (1044)[51].

2. Analyses

Dans cette section, nous procédons à une évaluation détaillée des résultats obtenus à partir des trois modèles et cherchons à comprendre les facteurs sous-jacents à ces différences. En outre, nous illustrerons certains cas d'erreur et les examinerons en détail.

2.1. Analyses sur les résultats des trois modèles

2.1.1. DistilBERT et BERT

Nous pouvons observer les performances du modèle DistilBERT sur l'ensemble de validation et de test à partir des tableaux 3 et 4 ci-dessus.

Au cours de l'entraînement, la perte (*loss*) du modèle DistilBERT diminue progressivement, ce qui indique que le modèle s'adapte progressivement aux données d'entraînement. La perte sur l'ensemble de test diminue également, ce qui suggère que le modèle se comporte bien sur les données non vues préalablement. Les indicateurs précision et rappel (*recall*) augmentent à chaque époque, en particulier sur l'ensemble d'entraînement. Cela signifie que l'apprentissage se déroule bien et que le modèle devient plus précis dans la reconnaissance de la classe positive des oronymes et qu'il rappelle également davantage de ces instances positives.

Le score F1, qui est une mesure combinée de la précision et du rappel, augmente également à chaque époque, ce qui suggère que les performances globales du modèle s'améliorent progressivement. En ce qui concerne l'exactitude, le modèle DistilBERT présente des taux élevés sur tous les deux ensembles de données, dépassant tous deux 0,97. Normalement, cela signifie que le modèle est capable de classifier correctement les instances positives et négatives dans le texte. Mais en fait, la raison pour laquelle l'exactitude est élevée est que la grande majorité des tokens ne sont pas des antonymes. Pour nous, c'est la F1 qui est important.

[48] Voir Annexe n°3 à la page 97
[49] Voir Annexe n°4 à la page 111
[50] Voir Annexe n°5 à la page 119
[51] Voir Annexe n°2 à la page 90

Dans l'ensemble, le modèle DistilBERT obtient de bons résultats pour cette tâche. Bien qu'il puisse présenter des performances légèrement inférieures sur l'ensemble de tests par rapport à celles sur l'ensemble de validation.

Bien que ces indicateurs puissent donner l'impression que les performances sont excellentes, les résultats que nous avons obtenus sur le corpus de validation réel, le `"list_without_oronymes"`, ne sont pas aussi satisfaisants. Comme mentionné précédemment, nous avons obtenu un total de 63 308 résultats, mais ceux-ci sont accompagnés d'un bruit considérable, notamment en raison de l'apparition fréquente des balises [SEP] et [CLS]. Leur occurrence atteint un chiffre élevé de 62 718, et dans la plupart des cas, ils apparaissent isolément. Il convient de noter que notre sous-corpus ne comporte que 63308 phrases au total. Cette situation suggère que le modèle a effectué des prédictions incorrectes.

Après une analyse approfondie, nous en sommes arrivés à la conclusion que cette situation pourrait être due au fait que nos données d'entraînement ne contenaient que des phrases contenant des oronymes. Par conséquent, le modèle semble avoir acquis une notion que chaque phrase dans le sous-corpus devrait contenir au moins un oronyme. Dans les cas où le modèle ne parvient pas à identifier un token d'oronyme réel dans une phrase, il pourrait choisir de marquer les balises spéciales [SEP] ou [CLS] comme faisant partie de l'oronyme.

Face à cela, nous avons décidé de retester nos données à la fois avec spaCy et BERT. Étant donné que le processus de résultat obtenu avec spaCy diffère de celui avec BERT, nous allons d'abord comparer les modèles DistilBERT et BERT, qui ont une relation plus étroite.

Comme nous l'avons présenté dans les chapitres précédents, nous avons filtré 1000 phrases issues des résultats de DistilBERT qui avaient été incorrectement prédites comme contenant des oronymes, soit celles comportant une balise [SEP] en isolation, et qui en réalité ne contenaient pas d'oronyme. Nous avons ensuite entraîné ces phrases sur BERT en suivant les mêmes étapes que précédemment. Comme indiqué dans les tableaux 6 et 7, les performances du modèle sur les ensembles de validation et de test sont excellentes. Dans la plupart des cas, les indicateurs de performance sur l'ensemble de test sont même supérieurs à ceux de l'ensemble de validation. Plus important encore, toutes les performances surpassent celles de DistilBERT.

Cependant, après avoir obtenu des résultats sur le nouveau sous-corpus, nous avons encore observé la présence des balises [SEP] et [CLS], bien que leur nombre ait considérablement diminué pour atteindre 2590, soit une réduction de 59 588 occurrences par rapport à la situation précédente. Le nombre total de résultats a également diminué de manière

significative, soit une réduction de 39 656 occurrences par rapport à nos observations initiales. Ces résultats préliminaires indiquent que l'ajout de 1000 phrases ne contenant pas d'oronyme dans les données d'entraînement réduit considérablement la probabilité d'erreurs lors des prédictions du modèle.

2.1.2. spaCy

Quant à spaCy, nous n'avons pas procédé à son entraînement avec notre corpus de données, mais avons directement utilisé le même sous-corpus que celui utilisé lors de la validation de DistilBERT. De plus, nous n'avons pas évalué ses performances en utilisant les mêmes métriques que pour les modèles DistilBERT et BERT. Nous avons obtenu un total de 83273 résultats, et sans surprise, ces résultats contiennent également une quantité significative de bruit. Le nombre de résultats obtenus est beaucoup plus élevé que pour les deux modèles précédents. Cela s'explique par le fait que nous n'avons pas restreint les résultats aux prédictions étiquetées comme étant de type "LOC", contrairement à ce que nous avons fait pour DistilBERT et BERT. Étant donné que spaCy est un modèle plus général, formé sur une variété de textes, il peut parfois manquer de précision dans des tâches spécifiques ou des domaines particuliers, ce qui peut entraîner des prédictions incorrectes.

Comme illustré dans le tableau 5, après le nettoyage des résultats, nous constatons que le label qui contient le plus grand nombre d'occurrences d'oronymes est "ORG", et non pas "LOC". De plus, plusieurs autres étiquettes telles que "PERSONNE", "FAC" et "GPE" contiennent également un nombre significatif d'oronymes.

En comparaison, DistilBERT et BERT sont des modèles d'apprentissage en profondeur spécifiquement conçus pour les tâches de compréhension linguistique. Entraînés sur des ensembles de données spécifiques à un domaine, ils sont mieux à même de capturer les schémas linguistiques pertinents pour ces tâches particulières. Lors des prédictions, ils tendent à focaliser leur attention sur les informations spécifiques à la tâche, ce qui réduit le risque de prédictions non pertinentes.

Cette analyse met en évidence les différences entre spaCy et les modèles DistilBERT et BERT. Alors que spaCy est plus généraliste et peut donner des résultats diversifiés dans différentes catégories, DistilBERT et BERT, en raison de leur entraînement spécifique, ont tendance à mieux performer dans des tâches spécifiques et à générer des prédictions plus ciblées.

2.2. Analyses des erreurs

Dans la poursuite de notre évaluation, nous examinerons des exemples spécifiques d'erreurs commises par chaque modèle. En analysant ces erreurs, nous espérons identifier les facteurs qui ont pu les provoquer. Cette approche nous permettra de mieux comprendre les forces et les faiblesses de chaque modèle et nous fournira des informations précieuses pour améliorer leurs performances futures.

Pour commencer, prenons l'exemple de "the Glacier d'Orny". Tous nos trois modèles ont réussi à identifier cet oronyme dans les phrases de notre ensemble de validation. Examinons maintenant les différences entre les résultats générés par ces trois modèles. Pour DistilBERT, le résultat est comme suit :

```
B-LOC: the, I-LOC: glacier, I-LOC: d, I-LOC: ', I-LOC: or, I-LOC:
##ny, B-LOC: [SEP]
```

Nous pouvons constater que le token "Orny" a été divisé en deux sous-lexicaux lors du traitement par le modèle. Chaque partie de l'oronyme a été correctement identifiée, mais le seul inconvénient est la présence d'une étiquette [SEP] supplémentaire identifiée comme le début de l'oronyme. En revanche, BERT nous fournit un résultat parfait, sans étiquette spéciale superflue. Le résultat de spaCy attribue correctement l'étiquette LOC à cet oronyme comme illustré dans la figure 4.

La principale raison derrière ces améliorations de BERT a déjà été expliquée dans la section précédente, résultant de nos ajustements sur les données d'entraînement.

Figure 11. Illustration de prédiction d'oronyme par spaCy (1)

Continuons avec le deuxième exemple qui illustre à nouveau une situation où les trois modèles ont des performances différentes. Les deux phrases suivantes tirées du sous-corpus contiennent l'oronyme "Aiguilles Dorées" sous ses deux formes en français et en anglais :

"Professor Forbes, as Balmat told me, named them, very happily, 'Les Aiguilles Dorées,' and they constitute the great feature of the pass."[52]

[52] Traduction : Professeur Forbes, comme me l'a dit Balmat, les a nommées, avec un grand plaisir, "Les Aiguilles Dorées", et elles constituent la grande particularité du col.

"They pour down from every break in the Aiguilles Dorées, as well as from the huge snow-capped heights on the opposite side of the glacier."[53]

Dans ce cas, seul spaCy a réussi à identifier ces deux formes, bien que les prédictions de balises soient incorrectes : "Les Aiguilles Dorées" a été identifié à tort comme "WORK_OF_ART", tandis que "the Aiguilles Dorées" a été identifié à tort comme "ORG". En revanche, ni DistilBERT ni BERT n'ont pu reconnaître ces deux formes, malgré le fait que la majorité des occurrences de l'oronyme dans les données d'entraînement commence par "the".

Après avoir examiné la liste initiale des oronymes utilisée pour l'entraînement, nous avons constaté que parmi les 15 occurrences contenant l'instance "Aiguille", une seule était au pluriel "Col des Aiguilles d'Arve", tandis que les 14 autres étaient au singulier sans aucun déterminant. Cette absence de variations dans les données d'entraînement explique pourquoi DistilBERT et BERT n'ont pas réussi à identifier les deux formes de "Aiguilles Dorées". SpaCy, grâce à son entraînement et à son optimisation sur une grande variété de textes, réussit à donner de bonnes performances dans de nombreuses situations, bien qu'il puisse également commettre des erreurs d'étiquetage.

Cet exemple met en avant les limites des modèles ainsi que leurs domaines d'application, soulignant la pertinence de l'utilisation de modèles généraux (tels que spaCy) dans des tâches spécifiques. En même temps, il met également en évidence l'importance des données d'entraînement et la nécessité d'effectuer davantage de prétraitement et d'ajustements pour s'adapter à des tâches spécifiques.

Prenons ensuite un nouvel exemple, avec la phrase suivante :

"A few minutes placed us on the pinnacle of the Bec du Mont Forchu, where we had, at least, a fine view of the Invergnuon."[54]

Nous devons identifier l'oronyme "the Bec du Mont Forchu". Voici les performances des trois modèles :

DistilBERT n'a pas réussi à identifier correctement l'oronyme complet, manquant la partie "chu" de "Forchu".

```
B-LOC: the, I-LOC: be, I-LOC: ##c, I-LOC: du, I-LOC: mont, I-LOC: for
```

[53] Traduction : Elles se déversent de toutes les brèches des Aiguilles Dorées, ainsi que des immenses hauteurs enneigées sur le côté opposé du glacier.
[54] Traduction : Quelques minutes nous ont placés sur le sommet du Bec du Mont Forchu, où nous avions, au moins, une belle vue sur l'Invergnuon.

BERT a réussi à identifier tous les tokens correctement, et l'oronyme a été reconnu de manière précise comme illustré dans la figure 5 obtenue à l'aide de l'outil de visualisation sur le site d'Hugging Face.

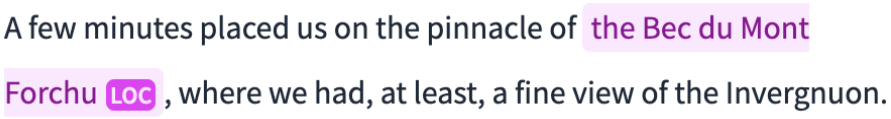


Figure 12. Illustration de prédiction d'oronyme par BERT

Cependant, cette fois-ci, spaCy n'a pas réussi à reconnaître du tout cet oronyme. Ce résultat est assez surprenant, car DistilBERT et BERT utilisent le même tokeniseur avec la même version. En ce qui concerne spaCy, la raison pour laquelle il n'a pas pu reconnaître cet oronyme pourrait être due à sa nature de modèle général. C'est l'une de ses limitations, car il n'a pas été optimisé spécifiquement pour la tâche en question et ne peut pas prendre en compte tous les oronymes envisageables.

Passons maintenant à un dernier exemple, qui mettra en évidence le cas où d'autres types d'entités nommées sont incorrectement identifiés comme des oronymes. Par exemple, dans la phrase suivante :

"My companions fortified themselves against the cold with kirschwasser;"[55]

Il n'y a normalement aucun oronyme, cependant nos trois modèles ont tous identifié le token "kirschwasser". DistilBERT et BERT considèrent qu'il s'agit d'un oronyme, tandis que spaCy a une opinion différente en le classifiant comme PERSON.

Il est également possible que spaCy considère un oronyme comme une personne, comme le montre l'exemple ci-dessous.

Figure 13. Illustration de prédiction d'oronyme par spaCy (2)

"Kirschwasser" signifie en allemand une boisson alcoolisée à base de cerises. La raison de ces résultats divergents pourrait être la présence, dans les données d'entraînement pour DistilBERT et BERT, d'une partie des oronymes en allemand. Quant à spaCy, cela pourrait être

[55] Traduction : Mes compagnons se sont fortifiés contre le froid avec du kirschwasser ;

dû au manque d'exemples similaires dans les données d'entraînement, ce qui l'a amené à faire une mauvaise interprétation en prenant "kirschwasser" pour un nom de personne en allemand.

Chapitre 4. Perspectives et réflexions

Dans le cadre de ce stage, nous avons constitué un jeu de données d'entraînement en projetant les 1221 oronymes à notre corpus. Nous l'avons utilisé pour optimiser deux modèles, DistilBERT et BERT, nous l'avons également testé sur spaCy. Toutefois, les résultats obtenus avec ces trois modèles différents présentent un niveau élevé de bruit et ont nécessité un nettoyage manuel approfondi. Bien que cette phase de nettoyage ait été longue, elle s'est avérée indispensable pour obtenir des résultats fiables et exploitables.

Avec le recul, plusieurs améliorations pourraient être apportées. En premier lieu, il serait recommandé de normaliser le texte dans le corpus dès le départ, car le corpus n'a pas été complètement normalisé. Comme nous l'avons mentionné dans le tableau 1, seules les années 1858, 1860, 1864, 1866, 1867, 1870 et 1872 contiennent des fichiers qui ont été vérifiés manuellement. Les autres fichiers contiennent des caractères spéciaux, ce sont généralement les résultats erronés d'OCR, ce qui pourrait affecter la précision des prédictions du modèle, voire entraîner des résultats incomplets. De plus, lors de la phase initiale de préparation des données d'entraînement pour le modèle DistilBERT, il serait judicieux de considérer l'intégration progressive d'exemples ne contenant pas d'oronymes. Par ailleurs, l'incorporation de nouveaux oronymes découverts dans les données d'apprentissage pourrait contribuer à enrichir davantage l'échantillon. Étant donné la diversité des oronymes disponibles, les données d'entraînement devraient refléter cette variété. Finalement, il serait conseillé d'opter dès le départ pour l'utilisation du serveur institutionnelle afin d'éviter les limitations de ressources computationnelles rencontrées avec des outils en ligne tels que Colab.

En ce qui concerne les compétences que j'ai acquises au cours de ce stage, j'ai développé une compréhension approfondie du processus d'entraînement de modèles de classification de tokens en utilisant des approches telles que BERT. J'ai assimilé l'impact de divers facteurs sur les performances du modèle, notamment la préparation des données, la sélection du modèle et la configuration des hyperparamètres. En outre, mes compétences en programmation Python ont été améliorées, ce qui m'a permis de manipuler et d'analyser les données plus efficacement. J'ai également pris conscience des limites inhérentes à différents types de modèles dans les tâches de traitement de texte, en particulier en ce qui concerne la prédiction effective des résultats réels. Cela renforce ma reconnaissance du lien étroit entre le traitement du langage naturel et la linguistique, ainsi que la nécessité d'une recherche approfondie pour soutenir leur

développement continu. De plus, en tant qu'enthousiaste de la géographie, j'ai enrichi mes connaissances sur les caractéristiques géographiques des Alpes.

Cependant, je reconnais que je dois améliorer mes compétences en ce qui concerne l'analyse des données et des résultats, en particulier lorsqu'il s'agit de grands ensembles de données. En outre, j'ai reconnu la nécessité d'acquérir plus d'expérience dans l'ajustement des paramètres des modèles. Dans l'ensemble, ce stage a été une occasion précieuse pour moi. Il m'a permis de progresser dans de nombreux domaines et d'acquérir une compréhension plus approfondie.

Conclusion

En conclusion, notre travail se concentre sur la reconnaissance automatique des oronymes dans un corpus de récits d'explorations alpines. Nous avons exploré l'utilisation de différents modèles de traitement du langage naturel pour résoudre cette tâche complexe. Nos résultats montrent des différences significatives dans les performances de DistilBERT, BERT et spaCy dans la tâche de la reconnaissance des oronymes.

DistilBERT et BERT réalisent des performances satisfaisantes dans l'ensemble, malgré le fait qu'il nécessite toujours un nettoyage manuel de la sortie. Ces résultats sont dus à leur excellente capacité intrinsèque à capturer des motifs linguistiques adaptés à des tâches spécifiques, ainsi qu'à l'entraînement sur des données spécialement préparées à cet effet. En ajustant les données d'entraînement, nous avons considérablement réduit les erreurs dans les résultats de BERT, ce qui souligne l'importance d'une sélection et d'une répartition appropriées des données pour assurer l'efficacité de ces modèles.

Cependant, spaCy, un modèle plus général, montre également des capacités impressionnantes dans la reconnaissance des oronymes. Toutefois, ses résultats sont moins cohérents, en partie à cause de son approche plus large de l'analyse linguistique. Bien que spaCy ait été en mesure d'identifier le plus grand nombre d'oronymes disponibles parmi les trois modèles, ceux-ci ont été classés dans diverses catégories d'entités, ce qui implique à nouveau un nettoyage manuel des résultats. En outre, il se trompe également dans la classification de certaines entités qui n'entrent pas dans la catégorie des oronymes. Cela souligne la nécessité de tenir compte du contexte et de la spécificité de la tâche lors de l'utilisation de modèles plus généralisés tels que spaCy, mais il peut servir de référence.

En résumé, notre travail met en évidence l'impact important des données d'entraînement et de la spécificité des modèles sur les performances de reconnaissance des oronymes. Malgré les inconvénients de DistilBERT et BERT, ces derniers constituent un meilleur choix pour des tâches spécifiques, tandis que spaCy illustre les défis posés par les modèles à usage général pour des tâches spécifiques.

À l'avenir, il est conseillé d'utiliser des modèles dérivés des Transformers, tels que BERT et DistilBERT, pour des tâches similaires à la nôtre impliquant la classification de token, en particulier lorsqu'un grand corpus doit être annoté.

Il est primordial est d'inclure dans les données d'entraînement une combinaison d’échantillons de vrais positifs et de vrais négatifs, tout en assurant un équilibre approprié entre

ces deux catégories. Ce ratio devrait être fidèle à la distribution des échantillons dans des contextes d'application concrets. Si une catégorie prédomine dans les données réelles, il convient d'intégrer davantage d'exemples de cette catégorie dans les données d'entraînement, ce qui permettra au modèle de mieux appréhender et saisir les spécificités de cette catégorie.

En outre, la diversité des données d'apprentissage peut contribuer à une meilleure compréhension des différentes situations et variations contextuelles, et par conséquent, à une amélioration de la qualité des prédictions. Cependant, il convient de noter que les modèles généralistes comme spaCy peuvent mieux convenir aux tâches de traitement de texte basiques et à certaines tâches d'extraction d'informations légères.

En fin de compte, notre travail contribue à une meilleure compréhension des performances et des limites des modèles de traitement du langage naturel dans la reconnaissance des oronymes, ouvrant la perspective d'explorer de meilleures méthodes de prétraitement des données et d'entraînement afin d'améliorer les performances des modèles et de mettre en œuvre des solutions pour que les modèles soient davantage adaptés à des tâches particulières.

Bibliographie

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). *TensorFlow : A system for large-scale machine learning* (arXiv:1605.08695). arXiv. https://doi.org/10.48550/arXiv.1605.08695

Bucila, C., Caruana, R., & Niculescu-Mizil, A. (2006). *Model compression*. *2006*, 535-541. https://doi.org/10.1145/1150402.1150464

Chollet, F. & others. (2015). *Keras*. https://keras.io

Comité français de cartographie. (1990). Glossaire de cartographie. *Bulletin du Comité français de cartographie*, *123, 124*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), Article 7825. https://doi.org/10.1038/s41586-020-2649-2

Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., Boyle, R., Cantin, P., Chao, C., Clark, C., Coriell, J., Daley, M., Dau, M., Dean, J., Gelb, B., … Ross, J. (2017). *In-Datacenter Performance Analysis of a Tensor Processing Unit*. https://arxiv.org/pdf/1704.04760.pdf

Kraif, O., & Diwersy, S. (2012). Le Lexicoscope : Un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques (The Lexicoscope : an integrated tool for combinatoric profles observation and lexico-syntactic constructs extraction) [in French]. *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, 399-406. https://aclanthology.org/F12-2033

Loper, E., & Bird, S. (2002). *NLTK : The Natural Language Toolkit* (arXiv:cs/0205028). arXiv. http://arxiv.org/abs/cs/0205028

Montani, I., Honnibal, M., Honnibal, M., Boyd, A., Van Landeghem, S., Peters, H., McCann, P. O., Geovedi, J., O'Regan, J., Samsonov, M., De Kok, D., Orosz, G., Blättermann, M., Altinok, D., Madeesh Kannan, Mitsch, R., Kristiansen, S. L., Edward, Lj Miranda, … Schero1994. (2023). *explosion/spaCy : V3.6.1: Support for Pydantic v2, find-function CLI and more* (v3.6.1) [Logiciel]. Zenodo. https://doi.org/10.5281/ZENODO.1212303

Ounoughi, S. (2023). The case of oronyms : Referential conventions as compromise. In L. Gardelle, L. Vincent-Durroux, & H. Vinckel-Roisin (Éds.), *Reference : From conventions to pragmatics*. John Benjamins Publishing Company. https://doi.org/10.1075/slcs.228.12oun

Ounoughi, S. (2024). « Une linguistique holistique du nom propre est-elle possible ? » *Corella*.

Ramshaw, L. A., & Marcus, M. P. (1995). *Text Chunking using Transformation-Based Learning* (arXiv:cmp-lg/9505040). arXiv. http://arxiv.org/abs/cmp-lg/9505040

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT : Smaller, faster, cheaper and lighter* (arXiv:1910.01108). arXiv. https://doi.org/10.48550/arXiv.1910.01108

Sennrich, R., Haddow, B., & Birch, A. (2016). *Neural Machine Translation of Rare Words with Subword Units* (arXiv:1508.07909). arXiv. https://doi.org/10.48550/arXiv.1508.07909

Stephen, G. (2004). The Alpine Journal : A century and a half of mountaineering history. *The Himalayan Journal*, *60*. https://www.himalayanclub.org/hj/60/1/the-alpine-journal-a-century-and-a-half-of-mountaineering-history/

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to Sequence Learning with Neural Networks* (arXiv:1409.3215). arXiv. http://arxiv.org/abs/1409.3215

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). *GLUE : A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding* (arXiv:1804.07461). arXiv. http://arxiv.org/abs/1804.07461

Williams, A., Nangia, N., & Bowman, S. R. (2018). *A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference* (arXiv:1704.05426). arXiv. http://arxiv.org/abs/1704.05426

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). *Aligning Books and Movies : Towards Story-like Visual Explanations by Watching Movies and Reading Books* (arXiv:1506.06724). arXiv. http://arxiv.org/abs/1506.06724

Sitographie

http://phraseotext.univ-grenoble-alpes.fr/Lexicoscope_2.0/

https://cercog.univ-grenoble-alpes.fr/

https://lidilem.univ-grenoble-alpes.fr/

http://www.alpine-club.org.uk/

https://demarreshs.hypotheses.org/

https://tei-c.org/ns/1.0

https://www.statmt.org/wmt14/translation-task.html

https://huggingface.co/

https://www.nltk.org/

https://jsonlines.org/

https://www.json.org/json-en.html

https://colab.research.google.com/

https://jupyter.org/

https://spacy.io/

Glossaire

Oronyme : nom porté par un élément de relief : une montagne, une chaîne de montagnes, une colline, une montagne sous-marine, etc. (Cartographie, Géographie, Toponymie) (Comité français de cartographie, 1990)

Toponyme : est un nom de lieu ; Source : https://dictionnaire.lerobert.com/guide/toponymes

Transformer : est une architecture de réseau neuronal conçue par une équipe de chercheurs de Google et de l'Université de Toronto en 2017.

Git : un système de contrôle de version distribué utilisé pour suivre et gérer les changements du code logiciel, permettant la collaboration entre plusieurs personnes, etc. Source : https://git-scm.com/

Grep: un outil de ligne de commande utilisé à l'origine sur le système d'exploitation Unix, écrit pour la première fois par Ken Thompson.

Fine-tuning : en apprentissage automatique, désigne l'ajustement final d'un modèle pré-entraîné sur un ensemble de données spécifique pour améliorer sa performance.

Taux d'apprentissage : un paramètre qui détermine la taille des pas que l'algorithme de descente de gradient effectue lors de la mise à jour des poids du modèle.

Régularisation : sert à prévenir le surapprentissage (*overfitting*) d'un modèle.

Surapprentissage : désigne une situation où un modèle d'apprentissage automatique s'adapte trop bien aux données d'entraînement, au détriment de sa capacité à généraliser à de nouvelles données.

EarlyStopping : est une technique de régularisation visant à prévenir le surapprentissage.

Taux d'apprentissage : un paramètre qui détermine la taille des pas que l'algorithme de descente de gradient effectue lors de la mise à jour des poids du modèle.

Régularisation : sert à prévenir le surapprentissage (overfitting) d'un modèle.

Batch size : (taille de lot), ce qui correspond au nombre d'exemples traités simultanément lors de l'apprentissage.

EarlyStopping : est une technique de régularisation visant à prévenir le surapprentissage.

Sigles et abréviations utilisés

AJ :	The Alpine Journal
NOMINALP :	Name of mountains in the Alps
OCR :	Optical character recognition
GPU :	Graphic Processing Unit
RNN :	Recurrent Neural Network
CNN :	Convolutional Neural Network
BERT :	Bidirectional Encoder Representations from Transformers
TAL :	Traitement automatique du langage naturel
NLP :	Natural Language Processing
MLM:	Masked Language Model
NSP:	Next Sentence Prediction
NER:	Named Entity Recognition
DistilBERT:	Distilled Bidirectional Encoder Representations from Transformers
GLUE :	General Language Understanding Evaluation
MultiNLI :	Multi-Genre Natural Language Inference
UNIX :	Uniplexed Information and Computing Service
NLTK :	Natural Language Toolkit; Source : https://www.nltk.org/
JSON :	JavaScript Object Notation
JSONL :	JavaScript Object Notation Lines
TPU :	Tensor Processing Unit
SHS :	Sciences humaines et sociales
TEI :	Text Encoding Initiative
XML :	Extensible Markup Language

Table des illustrations

Figures :

Tableaux :

Table des annexes

Annexe 1
Liste d'oronymes initiale

the Tête noire
the Col de Balme
the Glacier du Tour
the Tour glacier
the Glacier de Trient
the Grand Plateau
the Glacier de Taconnay
the Grand Mulets
the Glacier des Bossons
the Glacier de Salena
the Rigi
the Sidelhorn
the Oberaar Glacier
thethe Oberaar Col
the Gries Pass
the Grimsel
the Unteraar Glacier
the Dent de Morcle
the Nest Horn
the Monte Rosa
the Macugnaga Glacier
the Aiguille Verte
the Aiguille du Dru
the Glockner
the Pasterze Glacier
the Peak of Darkness
the Schreckhorn
the Great Schreckhorn
the Finsteraarhorn
the Finsteraar-horn
the Finsteraar Horn
the Finsteraar Glacier
the Finsteraar
the Lauteraar branch
the Zäsenberg
the Aar Glacier
the Strahleck
the Aletsch glacier
the Cima di Jazi
the Lötsch Thal
the Aletsch-horn
the Æggisch-horn
the Wengern Alp
the Lauber-horn
the Mönch
the Trugberg
the Alps
the Mer de Glace
the Gorner Glacier
the Saanenthal
the Wildhorn
the Sanetsch pass
the Fenêtre de Salena
the Wildstrubel
the Gemmi
the green Engstligen Thal
Entschligetal
the Niesen
the Räzli glacier
the Lämmerenhorn
the Simmenthal
the Amertenhorn
the Diablerets
the Olden horn
the Oldenhorn
The Bristenstock
the Valley ofthe Linth
the Linth valley
The Klausen
the Klausen Pass
the Balmwand
the Windgälle
the Clariden
the Gross Ruchen
the Scheerhorn
the Piz Rosein
the Bifertenstock
the Rüchi
The Klönthal Pass
the Klönthal
the Glärnisch
the Pragel
the Sand Grat
the Tödi Pass
rthe Sand Grat
the Kisten Grat
the Panix
the Segnes
the Segnes Pass
The Mürtschenstock
the Schild
the Fronalp
the Kerenzenberg
the Obere Sand Alp
the Glarus
the Catscharauls
the Rosein Alp
the Rusein Alp
the Blattenberg
the Sernft Thal
the Durna Thal
the Freiberge
the Kärpfstock
the valleys of the Serfnt
the Durna
the Linth
the Glärnisch
The Sardona
the Piz Valrhein
the Möschelhorn
the Splügen
the Scheibe
the Ringelkopf
The St. Peter 's Thal
the Savien Thal
the Calfeuser Thal
the Calanda-berg
the Aiguille du Moine
the Grande Jorasse
the Petite Jorasse
the Glacier du Léchaud
the Glacier du Géant
the Col du Géant
The Col du Geant
Le Rognon
the Aiguille Noire
the Alps
the Bosse du Dromadaire
the Dome du Gouté
the Col de Voza
the Aiguille du Gouté
the valley of Miage
the Col du Clou
the Croix de Feuillette
the Vélan
the Velan
the Grand Combin
the Col du Mont Rouge
the glacier of Corbassière
the Corbassière glacier
the Val de Bagnes
The Zmutt glacier
the glacier of Fenêtre
the Combin
the Passage de Graffeneire
the Graffeneire
the Avril
the Tour de Boussine
the Val d'Aosta

the Montagnes de Cogne
the Arolla
the Rouinette
the Mont Pleureur
the Matterhorn
the Great Matterhorn
the Cervin
the Truma des Boucs
the Crête Sèche
the glacier d'Otemma
the Petit Combin
the Combin de Corbassière
the Montagne des Cœurs
the Glacier de Mont Boveyre
the Glacier de Valsorey
the Gouille de la Vassue
the Gornergrat
the Triftbach
the Gnaphalium
the Rympfischhorn
the Col ofthe Adler
the Trift glacier
the Zinal glacier
the Gabelhorn
the Arpitetta
the Val d'Erin
the Val d'Anniviers
the old Weiss Thor Pass
the Cima di Jazi
the Nord End
the Gorner Glacier
The Adler Pass
the Riffel
the Old Weiss Thor
the Weissthor
the Höchste Spitz
the Breithorn
the Zumstein Spitz
the Lys Kamm
the Rosa
the Lys Glacier
the Grauenhaupt
the Zwillinge
the Cramont
the Weisshorn
the Dent Blanche
the Mischabel
the Aletschhorn
the Nesthorn
the Dom
the Täschhorn
the St. Nicholas valley
the Grabenhorn
The Mischabel horns
the Mischabel-Hörner
the Saas Grat
the valley of St. Nicholas
the valley of Saas
the Jungfrau
the AEggischhorn
The Laquinhorn
the Weissmies
the Trifthorn
the Bernese Alps
the Trift Alp
the Fletschhorn
the Adler Pass
the Allelein glacier
the Mischabel range
the Rympfischwäng
the Matterhorn
the Alleleinhorn
the Strahlhorn
the Weiss Thor
the Wetterhorn
the valley ofthe Adda
the Adda
the Bernina Pass
the Disgrazia
the Pointe des Ecrins
the Pointe des Arcines
the Crête de l'Encula
the Glacier Blanc
the Glacier Noir
Vallon de Sapinière
the Col de Selé
the Vallon de St. Pierre
the Pelvoux
the Crête de la Bérarde
the Val de St. Christophe
the Col delle Loccie
the Sesia glacier
the glacier of Zmutt
the Vispbach
the Viescher-grat
the Trugberg
the Mönch-joch
the Mönchjoch
the Eiger
the Faulhorn range
the valley of Grindelwald
the Jungfrau-joch
the Jungfrau Joch
the Twins
the Lyskamm
the Col des Jumeaux
the Glacier des Jumeaux
the Viescher-joch
the Oberland
the Viescherhörner
the Eiger-joch
the Lauinen-thor
the Théodule
the Faulberg
the Rothhorn
the Grande Casse
the Pourri
the Tarentaise Alp
the Klön Thal
the Findelen glacier
The Allelin Pass
the Adlerjoch
the Aletschhorn
the Alphubel-joch
the Great St. Bernard
the Ailefroide
the Glacier de la Pilatte
the Pilatte Glacier
the Glacier du Selé
the Col de la Tempe
the Écrins
the Aiguille du Midi de la Grave
the Glacier d'Arcines
the Balferinhorn
the Dent d'Hérens
the Val Vallante
the Viso
the Monte Rotondo
the Trugberg glacier
the Ochsenhorn
the Aletsch glacier
the Mettenberg
the Dru
the Glacier d'Argentière
the Argentière glacier
the Glacier de Triolet
the Tour Noire
the Chardonnet
the Mont Dolent
the Allée Blanche
the Col de la Tour Noire
the Oberaarhorn
the Grandes Rousses of Dauphiné
The Grandes Rousses

the Pic de l'Etendard
the gorge of La Cochette
the Pennine Alps
the Graian Alps
the Rateau
the Râteau
The valley of Allemont
The Range of the Meije
the Col des Cavalles
the Col de la Cavalle
the Glacier des Sausettes
the Glacier de Mousset
the Glacier de la Girauze
The Mont de Lans glacier
the Combe de Malaval
the Vallon de la Selle
the Grande Ruine
the Brèche de la Meije
the Col d'Erin
The Glaciers ofthe Bernina
the Bietschhorn Alp
the Blümlis Alp
The Studer Joch
The Dolomite Mountains
the Zufall Spitze
the Val Forno
the Sulden Thal
the Sulden
The Orteler Spitze
the Madatsch Glacier
the Monte Cristallo
The Video Spitze
the Confinale
the Vitelli Glacier
the Königsspitze
the Forno Glacier
The Viozzi Spitze
the Vios Spitze
the Val della Mare
the Palle della Mare
the Monte Tresero
the Zufall Spitze
the Val Furva
the Val di Sole
the Gavia Pass
the Val Bormina
the Orteler Group
the Laaser Thal
the Etsch
the Aussere Peder Spitze
the Schluder Spitze
the Weissmandl
the Vitelli Ridge
the Zebru
the Val del Zebru
the Confinale
the Monte Cristallo
the Oetz Thal
the Vorarlberg
the Lower Engadine
the Tresero
the Königs Joch
the Martell Thal
the Tabaretta Spitze
The Malser Heide
the Weisskugel
The valley of Martell
The valley of Ulten
The valley of Rabbi
The valley of della Mare
of Rendena
The valley di Genova
the Lötsch Sattel
the Jungfrau Joch
the Idateau
The Sesia Joch
the Moro
the Pizzo Bianco
the Val Sesia
the Parrot Spitze
the Sesia glacier
the Embours glacier
the Vincent Pyramide
the Weissthor
the Riffel
the Signal Kuppe
the Grindelwald
the Unteraar
The Enge
The Col du Mont Brulé
the Upper Grindelwald glacier
the Upper Grindelwald
the Tête Blanche
the Pic de Zardezan
the Col de la Traversette
the cockscomb
The Iséran
the Vaudet glacier
the Col d'Iséran
the Glacier des Eivittes
the Col Girard
the Glacier de Mulinet
the Vanoise
the Pointe de Séa
the Col de Séa
the Albaron
the Grande Motte
the Levanna
the Col de Collarin
The Col Tournanche
the Rothhorn
the Montanvert
the Col de Triolet
the Aig
the Talèfre
the Col de Pierre Joseph
the Petit Ferrex
the glaciers of Bagnes
the Alps of Cogne
the Tarentaise
the Pyramides Calcaires
the Col de la Seigne
the Bosse
the Dome
the Col de Miage
the Col du Chardonnet
the Glacier de Bionnassay
the Grands Mulets
the Grandes Jorasses
the Col des Grandes Jorasses
the valley of Lauterbrunnen
the Weisse Frau
the Roththal
The Maurienne
the Dent Parassée
the Glacier de Vanoise
the Lauterbrunnen
the Schwarz Mönch
the Silberhorn
the Giessen Glacier
the Schneehorn
the Guggi Glacier
the Grand Apparei
the Graians
the Col du Mont
the Val de Rhêmes
the Col Rosset
the Val Savaranche
the Glacier de Bassiac
the Col de Bassiac
the St. Hélène
the Croix de Nivolet
the Pennine Alps
the Aletschhorn
the Æggischhorn
the Ferpècle Glacier

the Orteler
the Ziller Thal
the Sulden Thal
the Heiligen drei Brunnen
the Gauli Pass
the Rosenhorn
the Berglistock
the Bergl
the Schneerinne
the Untere Schneerinne
the Obere Schneerinne
The upper Orteler Ferner
the upper Orteler Glacier
the Münsterthal
the Sassière
the Graian Alps
the Vaudet glacier
the Col Bassac
the Col de Gailletta
the Mt. Blanc range
the peak ofthe Grande Motte
the Col de Chavière
the Val Locana
the Valpelline
the Graignier de la Commune de Sixt
the Pointe de Salles
the Chaine des Fys
the range ofthe Anterne
the Col d'Anterne
the Col de Léchaud
the Buet
the Pic de Tinneverges
the Croix des Portes
the Gumihorn
the Fer à Cheval
the Méridienne
the Aiguille du Tour
the chain of Mont Blanc
the Besso
the Gabelhörner
the Dent Blanche
the Grand Cornier
the Schallenberg glacier
the Couvercle
the Col des Jorasses
the Col de Telleccio
the Piano di Telleccio
the Val Noaschetta
the Glacier du Vallonet
the Glacier d'Albaron
the Pte. de Chalanson
the Glacier du Collarin
the Alpe di Sea
the Col de Ciamarella
the Col de Sea
the Val de Tignes
the Val Groscavallo
The Todi
the Adula Gebirge
the Ober Sand Alp
the Altenoren Alp
the Pantenbrücke
the Altenoren Stock
the Clariden Grat
the Sand Alp
the Clariden Firn
the Piz Hussein
the Biferten Stock
the Eussein Alp
the Vals Cavrein
the Cavardiras
the Lukmanier
the Greina Pass
The Val Carassina
the Val Blegno
the Rheinwaldhorn
the Brenno
the Guferhorn
the Rheinwald Glacier
the Pena Alp
the valley ofthe Rhine
the Zapport Glacier
the Vallon des Etançons
the Roseg Glacier
The Laquin Joch
the Fee Gletscher Alp
the Laquin Glacier
The Laquinthal
The Bec de Luseney
The Ebnefluh Joch
the Schmadri Joch
the Agassiz Joch
The Viescher Joch
the Viescher Glacier
The Blümlis Alp
the Balmhorn
the Kastenstein
the Finsteraar Joch
the Agassizhorn
the Grünhorn Lücke
the Lotsch Thal
the Bell Alp
the Faulberg
the Tschingel Glacier
the Trift
the Nesthorn
the Jagihorn
the Bietschhorn
the Lotschenthal
the Mittaghorn
the Grosshorn
the aiguille of Argentière
the Great St. Bernard
the St.theodule
the Mont Blanc de Cheillon group
the Col de Colon
the Ruchen Glarnisch
the Maistein height
the Mittelgrat
the Ochsenjoch
the Kastenstein
the Berglistock
the Wetterhörner
the Schreckhörner
the Gspaltenhorn
the Renfer Joch
the Wetterhorn
the Lauteraar Glacier
the Miage Glacier
the Ruitor
the Aiguille de Trelatête
the Eis-meer
the Lauterbrunnen
the Staubbach
the Studerhorn
the Galibier
the Meije
the Jorasses
the Ailefroide
the valley ofthe Vénéon
the Col de la Lauze
the Glacier du Mont de Lans
the Combe de Malval
the valley of Chamounix
the cirque ofthe Viescherhörner
the Terglou
the Mangert
the gorge ofthe Savitza
the Kerma Pass
the Terglou Thor
the Drassberg
the Wurzen Save valley
the Kermathal

the Mischabeljoch
the Taschhorn
the Nadelhorn
the Domjoch
the Taschhorn
the Kien Glacier
the Jagerhorn
the Lyskamm
the Riffel
the Alphubel
the Fillar Glacier
the Schwartz Thor
the Betta Furka
the Jägerhorn
the Gorner Glacier
the Ludwigs-Hohe
the Feliks Joch
the Glacier of Felik
the Grand Paradis
the Grivola
the Poucets
the Tribulation glacier
the Rossa Viva
the Pic de la Lune
the Pic de la Tribulation
the Dzasset glacier
the Pointe de la Tribulation
the Ampezzo Pass
the Sexten Pass
the Auronzo Pass
the Etançons glacier
the Cavalles glacier
the Col de la Casse déserte
the Vallon de la Bonne Pierre
the Grande Aiguille
the Glacier de la Plate des Agneaux
the Roche Faurio
the Val Louise
the Roche d'Alvau
the Tête de la Charrière
the Bonne Pierre glacier
the Col de Galibier
the Brandenbergthal
the Crête du Glacier Blanc
the Glacier d'Arsine
the Sommet des Rouies
the Aiguille d'Olan
the Val Godemar
the Glacier du Chardon
the Mallet Glacier
the Klein Glockner
the Klein Silberhorn
the Klein Verra Glacier
the Val d'Ayas
the Brenta Alta
the Dolomite
The Rosengarten
the Brouillard Glacier
the Aiguille de Triole
the Dolent Glacier
the valley ofthe Arc
the low Col d'Arve
the Col de l'Infernet
the Comb
the Crête du Glacier Blanc
the Col du Glacier Blanc
the Vallon de la Pilatte
the Vallon du Chardon
the Glacier de la Casse Déserte
the Tête du Replat
the Glacier du Col
the Val di San Vito
the Corno del Doge
the Tre Croci pass
the Cristallo Group
The Gemerk
the Val Forame
the Val Grande
The Tre Croci Pass
the Col S. Angelo
the Val Popena bassa
the Crepo di Zumelles
The Forcella di Zumelles
The Grava di Spiolto
the Croda di Cesdellis
The Forcella di Pomagagnon
The Forcella di Bausa Marza
the Croda d'Orieto
The Cristall Pass
The Passo di Stauniès
The Rauhkofelschneide
the Schönleitenschneide
the Schönleitenpass
the Rauhkofelpass
the Forcella di forame
the Cresta biancha
the Grava di Stauniès
The Cadini Group
the Mesurina valley
the Rimbianco Alp
The Forcella alta del Monte Pian
the Höllensteinthal
The Forcella di Maraja
the upper valley ofthe Anziei
the Passo Campoduro
the Cadini
The Cadin della Neve
the Cadin di S. Lucan
the Forcella di Rimbianco
the Val di Rimbianco
the Forcella delle Biscie
the Forcella della Cima Cadino
the Fischleinthal
The Toblacher Riedl
the Paternkofl
the Toblinger Knoten
the Gwengalblspitz
the Paternkofl
the Rienzboden
the Neunerkofl
the Haunold
the Innerfeldthal
the Sextenthal
the Schwalbenkofl
The Innicher Riedl
the Dreischusterspitz
the Wildgrabenpass
the Pullkofl
the Bautkofl
the Steingebirge
the Hundstall Sattel
the Birkenkofljoch
the Höllensteinthal
the Dreischusterjoch
the Forcella di Giralba
the Val Giralba
the Kreutzberg
the Weiss Lahn
the Arzalbl Sattel
the Val Ambata
the Auronzo valley
the Forcella di Najarnola
the Passo di Piedo
the Drei Zinnen Joch
the Lange Alp
the Rienz
the S. Marko Knoten
the Sattel Berg
The Forcella di Lavaredo
the Santebüchl
the Patem Kofi
the Oberbacher Wand

the Altstein Thal
the Oberbacher Spitz
the Einer Kofl
the Oberbacher Alp
the Oberbacher Joch
the Oberbacher Thal
the Forcella di Giralba
the Forcella di Cengia
the Col Agnello
the Lämmerbüchl
the Forcella
the Santebüchl
the Lämmerbüchl
the Forcella del Col Agnello
the Forcella di Giralba bassa
the Val Riva
the Val Giralba
the Val Riva Secca
the Schreckhorn
the Lauteraar Sattel
the Passo del Basodano
the Tarentaise
the Glacier de la Gurre
The Graian Alps
the Pennine Alps
The Dauphiné Alps
the Glacier de la Selle
the Muzelle
the Ober Gabelhorn
the Zinal
the Col du Grand Cornier
the Dent d'Erin
the Arbengletscher
the Basodine
the Kastelhorn
the glaciers ofthe Adamello
the Primiero peaks
the Caré Alto
the Col des Aiguilles d'Arve
the Pic de la Muzelle
the Col de la Muzelle
the Zermatt glacier
the Tosa
the Gries
the lower Val Camonica
the Monte Campo Pass
the Tiefenbachjoch
the Pitzthal
the Tiefenbach Kogel
the Schwarze Schneide
the Taschach Glacier
the Pitzthaler Urkund
the Sexegerten glacier
the Wildspitze
the Hintere Brochkogl
the Hochvernagt Glacier
the Taschachjoch
the Hochvernagtwand
the Œlgrubenjoch
the Gepatschütte
the Kaunserthal
the Schwarzwandspitze
the Sulzthal
the Schrankogel
the Graba Alp
the Oestlicher Pfaff
the Zuckerhütl
the Sulzenauferner
the Wilder Pfaff
the Pfitscherjoch
the Sterzing
the Aiguille grise
the Lower Grindelwald
Glacier
the Mont Cenis
the Mont de la Saxe
the Val Ferrex
the Échaud
the Zillerthaler Ferner
the Thurnerkamp
the Gunkel-kaar
the Löffel Spitz
the Rother-kopf
the Zemmgrund
the Zemmthal
the Floiten Grund
the Ahren Thal
the Floiten glacier
the Mörchen Spitz
the Ross-ruck Spitz
the Horn Spitzen
the Rother-kopf
the Stillupgrund
the Tuxer Gebirge
the Croda Malcora group
the Marmarole group
the Antelao group
the Monti Meduce
the Marma-role range
the Forcella di Venodel
the Meduce
the Col Negro
the Val de' Bestioï
the Val di San Vito
the Forcella Grande
the Col del Prato di Mason
The Forcella del Val di Mezzo
the Punta del Val di Mezzo
the Forcella Piccola
the Antelao
the Forcella del Monte Bel Pra
the Cadore Pass
he Monte Casa Dio
the Cadin del Laudo Pass
the Pian di Begontina
the Selletta Pass
the Campo Marzo
Group
the Pian di Fraïnis
the Pian di Tardeba
the Begontina
the Punta Negra
the Croda della Cesta
the Monte Casa Dio Pass
the Cengia del Banco
the Tondi di Sorapis
the La Rotta Pass
the valley ofthe Boita
the Foppa di Mathia
the Val Vedessana
The Forcella di Marmarole
the Forcella di Saline
the Forcella di Peroi
the Forcella di Ruverde
the Forcella di Ciastelins
thethe Croda di Ciastelins
the Piave
the Forcella di Sondi Dove
the Val Vizza
the Forcella di Langerin
the Forcella di Palle
the Piano de' Buoi
the Forcella di Brusau
the Antelao Pass
the Forcella di Feltrume
the Campestrin Pass
the Cadin da Val
the Vallesina glen
the Croda di San Pietro
the Forcella Mandruzzi
the Forcella di Mandrini
the Forcella di Maisama
the Forcella di Nebbiu
the Forcella d'Onge
the Passo Campoduro
the Birkenkofl-Dreischuster

the Steinalbljoch
the Gsöllknoten
the Popera-Najarnola Group
the Eilferkofel
the Rothwandspitz
the Monte Popera
the Eilferkofeljoch
the Forcella del Val Riva
Secca
the Popera glacier
the Monte Cevedale
the Pizzo della Venezia
the Möselenock
the Mühlwalder
the Pusterthal
the Mühlwalder Thal
the Taufererthal
the Riese Gebirge
the Weissenbach
the Antholz
the Zemmgrund
the Hochfeiler
the Pfitscher Thal
the Mösele
the Furtschagel Ferner
the Schlegeisen Thal
the Gefrornes Wand
the Waxegg Alp
the Eiger Glacier
the Schneehorn
the Klein Silberhorn
the Eggischhorn
the Roththal Sattel
the Märjelen See
the Hoch Joch
the Brèche de Valsenestre
the Col de Lovitel
the Col de la Muzelle
the Glacier du Vallon
the Glacier du Montagnon
the Glacier des Cavales
the Col des Chamois
the Pic de Turbat
the Col d'Olan
the Pic d'Olan
the Glacier des Sellettes
the Col de Turbat
the Bec d'Invergnuon
the Bec de Glaçon
the Col de Grancrou
the Val Piantonetto
the Col de Tellecio
La Tour du Grand St. Pierre
the Col du Sonadon
the Col de la Maison Blanche
The Dent de Perroc
The Aiguille de la Za
the Col d'Hérens
the range ofthe Grandes
Dentes
the Col de Bertol
the Bec du Creton
the Val Pelline
the A. del Monte Moro
the New Weissthor
the Pic de Zinal
the Passo di Roffel
the Roffelstafel Alp
the Zinal Joch
the Glacier de Vigne
the Col delle Loccie
the Durand glacier
the Col Durand
the Gabelhorn Glacier
the Roc Noir
the Arben Joch
the Lüdwigshohe
the Pointe de Giordano
the Bors Alp
the Voudeck
the Sesia
the Ippolita Pass
the Lys Joch
the Glacier de Garstelet
the Glacier d'Indren
the Glacier d'Embourg
the Schwarzenberg glacier
the Bäregg
the Little Scheideck
the Seewinen glacier
the Faderhorn
the Eggfluh
the Tschingel
the Blümlis Alp
the Mittelegi
the Kalli glacier
the Gross Nesthorn
the Faulberg
the Bel Alp
the Ventina Glacier
the Schrötterhorn
the Schranspitze
the Königsspitze
the Ober Aletsch Gletscher
the Wenden Glacier
the Oberthal glacier
the Urathhorner
the G'schlotten See
the Wenden Joch
the Titlis
the Titlis Joch
the Passo di Ferro
the Val Bregaglia
the Val Bondasca
the Bondasca glacier
the Val Porcellizza
the Monte del Ferro
the Grigna
the Valle di Ferro
the Pizzo di Cocca
the Schrötter Joch
the Hoch Gall
the Hoch Wild Spitze
the Rainthal
the Defereggen Thal
the Krimmler Tauern
the Ahrenthal
the Fend Wild Spitze
the Langthaler Glacier
the Passo di San Lucano
the Wild Gall
the Schneebigenock
the Tofana
the Tre Sassi pass
the Val di San Lucano
the Primiero plateau
the Passo di Canale
the Coston di Miel
the Campo Boaro
the Passo delle Cornelle
the Palle di San Martino
the Palle di Rosetta
the Cima di Rosetta
the Lang Kofel
the Palle
The Sass Maor
the Cima di Ball
the Gran Sasso d'Italia
the Vette di Feltre
the valley of Primiero
the Abruzzi
the Rosengarten Gebirge
the Schallenberg
the Allerhöchste Spitze
the Marmolata
the Brenner Pass

the Grödner Thal
the Eisack valley
the Brenta Alta group
the Etsch Valley
the Seisser Alp
the Caressa Pass
the Kalbleck
the Rothewand
tve valley of Tiers
the valley of Karneid
the Falban Kogel
the Kessel Kogel
the Federer Kogel
the Rothewand Spitze
the Rosszähne
the Schlern
the Platt Kogel
the Latemar
the Sasso di Mugone
the Duron Thal
the Lausa Kogel
the Antermoja See
the Sella Spitze
the Rosengarten range
the Tierser Thal
the Oetz valley
the Stubay valley
the Cimon della Pala
the Cima di Vezzana
the Purgametsch
the Ross Zähne
the Schlern
the Schlern Bach
the Pania della Croce
the Alpi Apuane
the valley of Acereto
the Pisanino
the Cimone
the Pizzo d'Uccello
the Monte Sagro
the Apuan Alps
the Monte Altissimo
the Val del Monte
the Adige
the ParrotSpitze
the Ulia de Trieves
the Pierre Pointue
the Silberhörner
the southern Apparei
The Pennine chains
the Suldenthal
the Bergli-Stock
The Col de Bassac
the Ste. Hélène
the Fee Glacier
the Eismeer
the Kerma Thal
the valley of Auronzo
the Lavaredo Sattel
the Bühle
the Basodano
the Pfitscher Joch
the Zillerthal
the Löffel
the Riesige Patsch Ferner
the Pointe de Zinal
the Roffeljoch
the Glacier des Vigne
the Glacier de Piode
the Cornelle pass
the Rosszähne
the Pania
the Carrara Mountains
the Carrara chain
the Meduce
the Mer de Glace
Col Negro
John Ball
Grande Rousse
Pointe de Torrent
Zwolfer Kofl
Sattel der Pulle
Vecchio del Forame
Cima vanca
Val Blegno
Mer de Glace
Monte Rosa
Mont Blanc
Mont Mallet
Mont Tacul
Becca de Jazie
Castor
Pollux
Monte della Disgrazia
Monte Viso
Mont Dolent
Mont Suc
Mont Percé
Mont Broglia
Mont Fréty
St. Bernard
Mt. Iseran
Camadra Pass
Val Blegno
Val Soja
Piz Roseg
Mont Dolent
Milderstein
Mont Favre
Brèche du Râteau
Col de la Grande Ruine
Piz Popena
Monte Cristallo
Cristall Schönleiten
Punta del Forame
Val Popena bassa
Monte Campoduro
Boden Knoten
Monte Najarnola
Monte Castello
Monte Frerone
Wildspitze
Forcella di Venodel
Col Negro
Val Faloria
Croda Rotta
Valderino,
Val Salega
Vallazza
Val Sandoles
Val Langerin
Col di Lozzo
Monte Chiadin
Monte Tranego
Mont Durand
Monte Cevedale
Monte della Redorta
Mount Ventoux
Rusein Alp
Kalfeusen
Galanda
Dent du Midi
Cœur Signal
Einfisch Thal
Col Imseng
Le Peigne
Cavale
Val de Rhemes
Rusein Alp

Annexe 2
Intersection des données entre les résultats des 3 modèles

the col de luseney
the val del leno
glacier de la grande
the col de mesoncles
the aiguille du soreiller
the val sorapis
the glacier de vaudet
the col de la neuvaz
the val travignolo
the yaleille glacier
the weiten alp
mont de lans
the pointe de montandayne
val calanca
the col de fenstre
glacier de la casse deserte
val d'algone
the olden alp
the grohmann glacier
mont brule
pizzo campo tenca
the richetli pass
the val masino
the val toumanche
the grand coluret
the val de fournel
the zupo pass
becca de ortton
the plattas alp
piz pisoc
the petit plateau
the alp gnof
the glacier de gai
the col de la galise
val marson
the col de verbier
the val scura
the glacier de la tribulation
the col du celar
the criner furke pass
hangendhorn
the alp di balme
the col vert
the val grosina
the vigne alp
the alpe du tour
mont-blanc
the combe di valleiglia
the glacier of ayas
col de claire
the piz palil
the glacier de durand
the val orsine
the val vermolera
the piz fliana
the hooker glacier
the val de lys
the mont thabor
val antabbia
col de lauteraar
the bernetsmatt alp
the glacier du grandcrou sud
the val di vitelli
st. elias
the val gordolasca
the vallone del roc
val sparlotsch
the col de la casse deserte
the wajtersfirren alp
the palil glacier
the nantillons glacier
the col de champex
the col de luisettes
the croix blanche
the pic central
val di lanzo
the maloya pass
the glacier de la breche
col de rochefort
the hartley spitze
the ried pass
the val camonica
the col du loup
monte amaro
piz yadret
col vieux
the col de la lune
val formazza
the glacier carra
col du chateau des dames
monte vito
the urbach thal
the monte bosa
the cima del caire cabret
monte fernazza
etzthaler ferner
the piz medel
the glacier de monei
the grand pic
the fellaria glacier
the glacier de bertol
the lampertsch alp
col des diablons
the festi glacier
the zocco pass
the valle perse
the allelein pass
the baltschieder glacier
the russein thal
the pointe de sengies
monte cristallina
the turtman glacier
the pic verdonne
the piz bernina
cima dei gelas
the aiguille du gofitfi
the piz languard
col des rousses
the tour st. andre
val canali
val vajoletti
the glacier de getroz
the bertol glacier
the erstfeld thal
the pic jocelme
the col de girardin
val narcanello
the monte pian
mont clapier
the corno bianco
the gabelhom glacier
the col de la lanze
the glacier de bassac
the alp di ferro
faulhorn
the col de collon
the col de pila
the val maira
the glacier de la charpoua
the oetzthaler ferner
the morteratsch glacier
the col des cavales
the pointe du pousset
the hufi glacier
the monte prese
the konig spitze

col de galambre
the val prato
la balme
mont herbetet
the val champey
val jouffrey
the cima de jazi
mont vinaigre
the col du piolet
col du lion
pizzo stella
the uschinen grat
the glacier de prou
val bajon
the schweiben alp
the broglio glacier
the yerva pass
val asinozza
the nantulon glacier
the val niiglia
the cresta agiuza
the col de garin
col de nivolet
aletsch pass
col de blancien
the bricolla alp
the piz urlaun
the lebendun pass
val cengia
zervreilerhom
the zardesan glacier
the pic du midi
the aiguille de berenger
the montagne des coeurs
lower glacier
val fassa
val lavaredo
the tuiber pass
bruneggjoch
the grandcrou glacier
col de valloire
the drei schuster spitze
the glarnisch
cima del pizzo
the punta del broglio
the bisi thal
the plachten alp
the herbetet
the geschenen thal
the gallo pass
fillar pass
val masino
gemshorn
vin du glacier
monte rosso
the alpe ravina
vallouise
the col des ecandies
the ziilerthaler ferner
the val de peychouda
monte kosa
the vallon de la sausse
the col de lauzon
the munster alp
the vogel joch
the col de vars
the val giuf
the punta giordani
the fox glacier
the col de sonadon
the glacier de marinet
the col de la breya
the col de la za
the darrei glacier
the presena glacier
the aiguille du plat
the col de la temple
the seisser alpe
the yentina glacier
the schwarz glacier
val anzasca
the col de raus
the bossons glacier
piz borel
the val tuoi
the fall glacier
the fenetre de cogne
the col del carro
the borzago glacier
piz bellavista
the hied glacier
the bezingi glacier
the scaletta glacier
the como bianco
col du grand sauvage
the tre coci pass
the pizzo san colombano
col budden
the trogen alp
the hangend glacier
the col de fenetre
val mora
val del sasso
the defereggen thai
spitzbergen
the viubez glacier
piz vadred
the pointe de mary
the col du tacul
the glacier de lauzon
monte cavallo
the wingern alp
piz vadret
the augstbord pass
val grisanche
the piz ner
monte civita
the bietsch joch
the monte perdido
the hanig alp
the vallon de la mariande
the hohberg pass
piz palii
the glacier de chauvet
the col vicentino
the val lavizzara
val di rabbi
the col des fours
the val de tinges
the col toumanche
aiguille de polset
the aroila glacier
birchfluh pass
val verzasca
val canzoi
the col de la grande luis
furggthal
the fiinffinger spitze
the col de vallon pierre
the val cairos
piz morteratsch
the col de yallante
the abberg glacier
the col de la gippiera
the col de la plate des agneaux
col de la fenetre
the aiguille blanche de peuteret
col du cormet
the sasaello pass
the col du tour
monte zigolo
piz roz
the telleccio glacier
the vallon de chillol
the col du says

the col de goleon
the pointe de rosablanche
the brenva glacier
monte foscagno
the zwischbergen pass
pic des agneaux
the becca di luseney
the piz sella
pointe de mary
piz zupo
the punta foura
mount st. elias
col de castelnau
the val campaccio
stellihorn
the kistengrat pass
the drakosh pass
val di fontane
the col de traversette
the aiguille du glacier
the rhone glacier
the hochste spitze
the mont de saxe
the val vecchia
pizzo venezia
serravalle
the col du bonhomme
the col du lion
piz chalchagn
the col di finestre
ochsenthal
the pre de bar glacier
vallante
the col des courtes
the col de mont rouge
bies joch
balmenhorn
the col du sellar
the val di forzo
the wengern scheideck
the col de la nouva
the montandeyne glacier
the hochbalm glacier
the col de cheillon
the rocher de naye
the col du talon
the misauna alp
the monte eosa
the fraele pass
the col de rochefort
the pointe de bricolla
the glacier d'invergnuon
the scaletta pass
the glacier de zardesan
the cristallo glacier
the fomo glacier
the piz eoseg
the col des masses
the sella pass
the val cavrein
piz medel
viescherhorn
the glacier de pabeille
the col de eochefort
the col des grandes murailles
the cima de jazzi
becca di monciair
aiguille du midi
the stein alpe
the val de rhymes
val belviso
the biferten glacier
the col de tignes
grohmann spitze
the montandevne glacier
the glacier des eivettes
ost spitze
mont albert
the vallon claus
the becca di nona
the col de blancien
piz st. michael
the val fassa
mittelhorn
the val del zebra
the col della piccola
the col de tracuit
the val de la gitte
the schwarze glacier
val rendena
the glacier carre
piz bernina
val lavizzara
pied du col
the glacier du giant
the pointe de chamossaire
glacier de la bonne pierre
the glacier de la bonne pierre
the monciair glacier
the col de la coste rouge
the felik joch
the gomerer thal
val di zoldo
the val lavaz
the glacier de lechaud
the glacier de freboutzie
the gredetsch glacier
the glacier de parste
the pic tyndall
the val peisey
val travignolo
flavona alp
val malenco
the pizzo tremoggia
the columbus glacier
the bemetsmatt alp
monte sissone
mont ruan
col de la lune
the gauli glacier
the col de galese
the col de rhymes
the glacier du fond
the col de valsorey
val di scalve
the brunni glacier
the col dolent
the val de ribou
val tournanche
the amethystes glacier
the ried glacier
the charpoua glacier
the sehalliberg glacier
leghorn
the piz buin
the monte plessura
the col de jaman
felik joch
the col du tour noir
the grenz glacier
the glacier de tacconai
the val di gesso
the pointe de la sana
col de cretan
mont colon
the aiguille de la peau blanche
the glacier de rochefort
the glacier du grand mean
the col lombard
col du vallon
piz minger
monte folletto
the crusehetta pass
gebelhorn
the glacier de nant blanc
the glacier de charpoua

the croix de belledonne
valaisan
the col de mont brule
the glacier du brouillard
the glacier de galese
the colle budden
the thaileit spitz
the corno piccolo
the argentiere glacier
val bavona
the monte nero
the lenta glacier
val travernanzes
val brenta
the bresciana glacier
val nambino
the monte maurigno
gassispitz
val chigniulascio
the cimes blanches
the bresciana alp
val federia
piz linard
the col de giers
the zmutt ridge
val lavinuoz
val cabione
the col de sageroux
the vallon des bancs
the pic d'otemma
the roccia viva glacier
the tuoi glacier
val fiera
the cima viola
val cluozza
the pic olan
the col de bellevue
col de la croix
the gliems glacier

scerscen pass
furgen glacier
val vigornesso
the glacier de cijordnove
mont cheillon
bieshorn
the glacier du tabuchet
val di forzo
the padeon alp
the col de mont brute
the maderaner
col de chalance
the nagler spitz
fillar joch
the rhone valley
the aiguille de la yola
the glacier de yaudet
the pointe percee
linththal
val de lys
the val di valasco
piz quatervals
the pic de montandayne
the jung pass
the val jouffrey
val bendena
veglia alp
the aiguille du gouter
the ciardonei glacier
the tete blanche
the chardon glacier
the val del sasso
the antabbia glacier
the monte di zocca
rhone glacier
val ostera
the netz glacier
the yuibez glacier
the col de mary

the alpe granus
monte leone
the farno glacier
the col de galdse
the col de girarain
the aiguille de charmoz
the grodner joch
pic jocelme
colle budden
the dents des bouquetins
val devero
the zebles pass
the vallon de sellavieille
the val varaita
the hohberg glacier
the val di rabbi
the col de pargentine
the almagel alp
mont mounier
the val de hibou
the gepatsch glacier
the val bavona
the scerscen glacier
the val di roda
the col de creton
the col de la pilatte
the grand tour
the col del merlo
monte vibo
the col de vallante
the col de brouis
monte pian
the val des bans
col de puissaille
the glacier du grand
the herbetet glacier
the val grisanche
vallette
the val maisas

the vin du glacier
the piz humor
piz plavna
the glacier de grandcrou
the hohsand glacier
the st. elias
the val piantonello
the col de lautaret
the vincent pyramid
the gletscher alp
the zumstein spitze
yispthal
val pravitale
the col di cerieja
the dolomites
col du thabor
monte rotta
rinderhorn
the grande fourche
monte gristallo
the glacier de monestier
the col de la scigne
the staffel alp
val formin
the pic de la grave
val buona
the huddleston glacier
the maderaner thal
the glacier de cheillon
the glacier de za de zau
the val maggia
the medel glacier
the foppa alp
the piz tschierva
the tasch alp
the fillar alp
the pic de neige cordier
the silvretta glacier
the gross ruchi

the col du fond
the lauzon glacier
the breche glacier
the peuteret ridge
val teresenga
the montandayne glacier
mont rose
col de la muande
col du sirac
the salena glacier
the col di lago
the becca de nona
vispthal
the pioda alp
the pas de turloz
the col de la sauce
the blumlis alp
the mont perdu
the pic du glacier
the bee du grenier
the oberhom alp
monte bignone
the porchabella glacier
val leventina
the val zebra
the glacier de l'arpont
val antigorio
the schmadri glacier
fluchthorn
colle fiorito
the clos de la cavalle
the mesurina alp
val seriana
the col de la portetta
the dungel glacier
the ost spitze
arpette glen
val maggia
the vuibez glacier
piz roseo
the mestia pass
the val di boda
col des aravis
val noana
val orsine
the moncorve glacier
the sella joch
the corbassiere glacier
the rosenlaui glacier
the val de cogne
the col de la casse
piz buin
the cedeh glacier
the weingarten glacier
the col de champorcher
the moming pass
becca di nona
the est spitze
the col de sassa
the buffalora pass
mont joli
the ausser locker spitze
the wengera alp
the col de la muande
the cima cadino
val maisama
the fassa thai
the aiguille balmat
mount ararat
the col des hirondelles
the val de st. marcel
the drei zinnen
the trou de toro
brunegghorn
the val de grauson
sella pass
the alp oberkasern
the glacier de fos
the joch pass
the val malenco
the tabaretta thal
col des trois pointes
the becca di noaschetta
the glacier de grandcrou sud
val anziei
the buss alp
the col de lauteraar
the otemma glacier
the col du lauteret
pic verdonne
monte durano
the alp kobiei
the ban glacier
the saleinaz glacier
the noaschetta glacier
the upper brenva glacier
the pic du frene
val coumera
monte matto
the col bardoney
dent parassee
montenvers
the col de la maigna
the pointe de yaleille
the mont fleuri
the zagen glacier
the glacier de pierre joseph
the fedaja pass
gassispitzen
egginerhorn
val mergoscia
col des rouies
the val d'entremont
the piz terri
the val vraita
the pizzo valgrande
the val ferret
val di susa
the mur de la cote
the monte bignone
the col de nivollet
the tete noire
the col du lautaret
col des grandes bousses
the col de baline
the aiguille du plan
val joufirey
the pizzo del diavel
the alp zarmine
the brunni alp
the col de la pointe de bricolla
the glacier de blaitisre
the pic de retour
the clos des cavales
vallombrosa
the glacier de tronchey
the wengem alp
the furgge glacier
the chamois pass
the passo rovano
the estellette glacier
monte augusta
the val stura
col de valsorey
the happen glacier
the roseg thal
the val champagny
the val tournanche
monte oristallo
the aiguille de peteret
the val di vallante
rhein thal
col de rioufroid
the aiguille du midi
the col de la gailetta
the glacier de moiry

the col de monciair
piz murtarol
the gorner grat
the val tellina
val giuf
the val di braulio
val livigno
the roccia viva
the zapport alp
piz tavrii
the langthaler joch
piz palu
the hasli thal
the col de vauon pierre
the col du thabor
the col de la lavey
the verpeil spitze
becca conge
the nantillon glacier
the glacier des amethystes
the combe froide
the petit mont
the ampezzo valley
val yerzasca
montets
rizlihorn
the pic lory
val di mel
the col di telleccio
the fresnay glacier
the la neuvaz glacier
the val sinestra
the grande plaque
the cogne mountains
the plan de la cavalle
mont maudit
the crozlina alp
the cima bianca
the valle grande

the col bonney
the glacier du grand tetret
the val de la leisse
the parung pass
the tic du glacier
the turtmann glacier
the schalliberg glacier
the gietroz alp
pointe de sainte anne
the rocher blanc
the col de sais
the siidlenz spitz
the basso di muro
the mont pelat
the col du couard
the val calanca
col verdonne
the pujo glacier
the val carapiglia
the col de monei
the arbola glacier
the sulzenan glacier
the grindel alp
valetta
the grand colouret
val campedelle
the konigs spitze
the col du gdant
the piz morteratsch
the aiguille de blaitiere
mutthorn
the val di canale
the oberalp see
the col de frette
the col des aravis
the nuefelgiu pass
dolomites
the como piccolo
col de fenstre

the cavergno glacier
the pic de bure
monte caggio
the pointe des plines
aiguille de la sausse
monte gazza
the punta bianca
the grand glacier
the glacier de sengie
monte zovo
monte eosa
monte como
the glacier du mulinet
the pizzo columbe
the col cliamonin
monte gallina
the maya de bricolla
the col de martignare
tete noire
the mont colon
the petit coluret
allerhochste spitze
monte ventina
the val campiglia
the ginevrie alp
the glacier de la frasse
the pointe des salles
the col ferrex
faderjoch
col de la leisse
the grosse windgelle
the col du palet
the mont falcon
the pic gaspard
col becker
the glacier de broglia
saasthal
the vertain spitze
the col de grandcrou

the col du charforon
schallhorn
the corno del camoscio
the glacier des ignez
the val lavinuoz
the glacier du geant
the col de belleface
the tschierva glacier
the val des ormonds
the piz linard
the pointe de la yalettaj
the verra glacier
the rochers rouges
mont lacha
the petit mont colon
the col de berard
la tour ronde
the monch joch
pic central
the piz formin
the col de chalance
the aiguille de miage
the ampezzo dolomites
monte di pietit
the becchi della tribulazione
the col theodule
val imagna
banhorn
piz kesch
the col de yalpelline
monte bocchetta
the pointe du colloney
the val fiera
the sonadon glacier
pointe de marguerite
the col de la brenva
the cima del lago agnel
the val ciamoseretto
val grana

the corno di dosde
the flavona alp
the dossradond pass
valsorey
the pers glacier
the pic oriental
monte cistella alta
monte bego
monte mottarone
the val de galambre
the col de pevsque
val di prato
the monei glacier
monte nero
the col de la croix haute
the liappey alp
the pisole alp
mont emilius
the pic du thabor
piz formin
the allalin glacier
val di lucerna
the col de joux
monte morrone
the col di tenda
the becco di mezzodi
the fum glacier
the glacier de tabuchet
val ombretta
col de la sauce
the val fomo
val codera
glacier des ecrins
the cima di canale
the langthaler ferner
the glacier du casset
rocher rouge
val camonica
the val devero
the rochefort glacier
the tiefenmatten glacier
the noasehetta glacier
monte spinale
the bies glacier
the glacier de miage
glacier du clot des cavales
the val mora
the baltschieder joch
bardoney glacier
the gasteren thal
the val angrogna
the gran neiron glacier
monte blanc
the pont de mauvoisin
val fonda
tour noir
the monte oliveto
the saas thal
val vermiglio
the col du midi
the alp crozlina
doldenhorn
the col di teleccio
valsenestre
the nagles spitze
the glacier de la casse deserte
the schrotter joch
the pic bonvoism
the blumlis alp glacier
the col de jallorgues
the pizzo di ferro
valtellina
the pointe de garin
the matscher thai
the col de charnier
mont pelat
the pizzo campo tenca
val malvaglia
colle campaccio
sckontauf spitze
the tour st. pierre
the sissone glacier
glacier de thomme
the hochbalen glacier
the val lazin
augstbord pass
col emile pic
the allerhochste spitze
the monte rotta
the glacier de zardezan
the bee de phomme
the bocca dei camuzzi
the glacier du bouchet
the piz zupo
the val de viu
the col de bardoney
the glacier de saleinaz
monte eotta
the col du vol
monte forato
the glacier des nantillons
glacier de freboutzie
val challant
val ruvinian
the val pradidali
the val di mello
the val de pagnelle
the glacier des ignes
the pointe de tinneverges
the hussein thal
mont roselette
the bosco nero alp
the col vaudet
the dent de corjon
the cappella di monte
the teleccio glacier
the glacier de chaviere
the bies joch
the glacier de brenva

Annexe 3
Intersection des données entre les résultats de DistilBERT et de BERT

col du sele
the col de la neuvaz
the grunhorn
the valley of zmutt
the col de cheville
the grohmann glacier
the fletsch glacier
the val de fournel
the zupo pass
the mont clapier
the tinneverges
the palentina alp
the col de montand
the snowdon range
the col de la galise
alp gnof
the vigne alp
the combe di valleiglia
col de claire
the glacier de durand
the mont thabor
the baltsehieder joch
the lioththa
the val gordolasca
the rosenlaui
rosenlaui
the val di bin
the vallone del roc
the wajtersfirren alp
the pic central
val di lanzo
the langefluh
the ried pass
punta forches
the glacier carre fall
val formazza
the col de la lune
the monte bosa
the cima del caire cabret
the valdieri
the roche melon
the glacier de bertol
the val di cougne
the hiillehorn
the allelein pass
col des rousses
val canali
the erstfeld thal
the goleon
the tschierva morteratsch

glacier

the alp di ferro
the maglich ridge
the colie della piccola
pic de la grave
the col des cavales
the pointe du pousset
mont jovet
la balme
the col de la
monte rota
the dussistock
the col du piolet
the viescherhorn
the strahl-grat
col des sellettes
the charpoua
the val niiglia
the val di lys
the bricolla alp
the fradusta
the lebendun pass
val cengia
zervreilerhom
the dachspitze
the pic du midi
the aiguille de berenger
the montagne des coeurs
st. nicolas
the grandcrou glacier
the glarnisch
the grodener valley
the latelhorn
the rizlihorn
the gallo pass
vin du glacier
monte del castello
soureillan
the griesthal
the alpe ravina
the jungfraufirn
the richetli
the bothhorn
col della traversetta
the val de peychouda
monte kosa
the colle perduto
the schallbetalp
the glacier de piece
the val giuf
the lower kerma alp
the col de la breya
valiy
the aiguille du plat
the valley of hasli
val anzasca
the col de raus
the val tuoi
the fenetre de cogne
the vieschergrat
the hied glacier
thechste spitze
the diavolczza pass
the aiguille d'argentieres
col budden
the col de fenetre
the monte delle loccie
the saasthal
val del sasso
piz vadred
the pointe de mary
the col du tacul
the balenfimhorn
the montblanc
the alphubeljoch
the wingern alp
piz vadret
val grisanche
the combe di valeiglia
aiguille du geant
the glacier de lauzon
the glacier de chauvet
the combe di telleccio
val verzasca
the pic du re tour
the monte frappa
the col de vallon pierre
piz morteratsch
col du cormet
the col noaschetta
col lombard
the kandersteg
monte zigolo
the vallon de chillol
schrammacherspitze
the schonbuhl glacier

the parrotspitze
the kistengrat pass
the valley of chamonix
the drakosh pass
the bettenhorn
the rhone glacier
the montenvers
serravalle
piz chalchagn
col du grand tetret
the pelmo
the pian di sera
the monte eosa
the valtendra
the scaletta pass
the eigher
the cristallo glacier
the casera del becco
the ahrenthalers
the maderaner thai
viescherhorn
shone glacier
becca di monciair
aiguille du midi
the stein alpe
the miinsterthal
the biferten glacier
the rolle pass
grohmann spitze
the aiguille du geant
the amerten glacier
the mettelhorn
the col de la pi latte
mont albert
jung pass
the col de blancien
the val fassa
mittelhorn
col de petit pierre
the val de la gitte
the schwarze glacier
columbe
the fisistock
the erstfelderthal
the kleinalpthal
val lavizzara
pied du col
the garstelet glacier
the glacier de la bonne pierre
the alefroide
the monciair glacier
glacier de
schwarzhorn
the yalais
the val peisey
the roththalhorn
the valle di vih
col de la lune
the tour du grand st pierre
the sagis alp
the eisjochl
the lower trubsee alp
the amethystes glacier
the monte plessura
the hockhorn
the vallon du diable
the calfeisen
the glacier de tacconai
the glacier des eta
the pra fiori
the aiguille de la peau blanche
the valley of the souloise
gebelhorn
the glacier de nant blanc
the gasenthal
valaisan
the col de mont brule
the punta di cian
the fimberboden
the bresciana glacier
gassispitz
the aiguille du goute
val fiera
the pizzo venezia
the cenis
pic des posets
the rothgiatli pass
val vigornesso
hornli
the monchjoch
val di forzo
the padeon alp
the glacier de yaudet
the faulhorn
the tosenhorn
the jung pass
the schrund
the monte giralba
the bernina
the grand crou glacier
rhone glacier
val ostera
the netz glacier
the farno glacier
the col de galdse
the col de girarain
the aiguille de charmoz
the grodner joch
pic jocelme
the wildgrat
the dents des bouquetins
val devero
the buttlassen
the zebles pass
colle budden
the fluchthorn
the fletschalp
the pedriolo alp
the mattebhorn
the puster-thal
the col de la pilatte
the arpette glen
monte vibo
the col de brouis
the val des bans
the val di molini
the piz humor
piz plavna
the glacier de grandcrou
the hohsand glacier
the vincent pyramid
yispthal
the col di cerieja
the dolomites
the dome
the colouret
the gemshorn
the glacier de monestier
the aiguille du croissant
the bernina cone
the jungen glacier
the blinnenhorn
the tasch alp
the fillar alp
the pic de neige cordier
the gross ruchi
the lotschenliicke
the stubaithal
mont rose
the brenta bassa
the salena glacier
vispthal
piantonetto
the pioda alp
the aletsch
the glacier des rousses

the pas de turloz
novaliciense
the hohberg pass glacier
the val zebra
simelistock
fluchthorn
the triolet glacier
colle fiorito
the bietsch-horn
val maggia
the pala di s. martino
the col de pherbetet
the moncorve glacier
the barenthal
the albula
mont joli
the salena
the valette
moming pass
the etzlithal
mount ararat
the aarthal
the val de grauson
the alp oberkasern
the glacier de fos
the fiinffingerspitz
col des trois pointes
the becca di noaschetta
the glacier de grandcrou sud
col de la pilatte
the heiligenbergspitze
the col du lauteret
col des cavales
the pic du frene
the grande serre
the col de la maigna
the pointe de yaleille
the glacier de pierre joseph
gassispitzen
the lohnerhorn
the val rusecco
the vogelberg
the val ferret
damma pass
the pizzo porcellizzo
roches moutonnees
the col de nivollet
the tete noire
col des rouges
the jungthaljoch
vallombrosa
the wengem alp
marschollhorn
the estellette glacier
col de valsorey
the bilchlistock
col de sais
mont sonadon
the piz michel
the st. gothard
rhein thal
the glacier des etancjons
val livigno
the roccia viva
the zapport alp
the lysjoch argte
pic de la pyramide
the petit mont
val yerzasca
montets
san nicold
the hintergrat
the cima del gelas
the cime des torches
the valle grande
schneekuppe
the col de la selle
mont dolin
the col de la galese
the punta sengie
the turtmann glacier
picco della speranza
the gietroz alp
the canali
the siidlenz spitz
the glacier de sea
the simelistock
the col du couard
the zwolfer
the plan de la tribulation
the hornli
the col de seguret foran
valetta
the piantonetto
the col du gdant
mutthorn
the val di canale
the col des aravis
the huhnerstock
dolomites
the val della noana
col de fenstre
the valley of cogne
the grenzgipfel
the balm-horn
the ofenhorn
the pic de bure
the indren glacier
hochste spitze
the tour noir
aiguille de la sausse
the glacier des bois
the kamm glacier
the glacier du mulinet
the rocher
the col cliamonin
the pizzo columbe
tete noire
the mont colon
the val fedoz
the val campiglia
col de la za
the colle di livoumea
the grosse windgelle
saasthal
the vertain spitze
schallhorn
the cima di saoseo
the valley of arsine
the val des ormonds
berthemont
the oetzthal
the piz linard
the augstenberg
the thalihorn
mont lacha
the piz formin
the lauteraarhorn
berhorn
the col theodule
monte bocchetta
val grana
valsorey
the periades
the pic oriental
monte bego
monte mottarone
the groraer glacier
the col du'geant
the furggen glacier
the col de la croix haute
the pic du thabor
the allalin glacier
monte morrone
montrond
the colloney

the valley of maurevielle
the val fomo
col de la sauce
the cima di canale
rocher rouge
piz vadbet
the valeiglia
the saas thal
the sasso di muro
the col di teleccio
the nagles spitze
the glacier de la casse deserte
the schrotter joch
the pizzo di ferro
the bosetta
the triftjoch
the zinareffien
the tour st. pierre
the sissone glacier
the bernina glaciers
the asggischhorn
the nest glacier
the punta lazin
the æggisch-grat
the monte rotta
the bocca dei camuzzi
the piz zupo
the cian ridge
monte eotta
the glacier des nantillons
glacier de freboutzie
the rauhhofelpass
the val de pagnelle
the col du cret
the col vaudet
the glacier de chaviere
the tschingelhorn
the col de luseney
the val del leno

mont gele
the klonthal
the val sorapis
the glacier de vaudet
the val travignolo
the weiten alp
mont de lans
mont brule
pizzo campo tenca
the richetli pass
the eggerhorn
the langkofel
the plattas alp
the sefinen alp
the glacier de gai
the valley of aletsch
thevreilerhorn
novaliciens
the val scura
the glacier de la dent
the windgelle
hangendhorn
the dognathal
the val grosina
the hintere schwarze
tabel glacier
mont-blanc
the glacier of ayas
the grandes dents
the piz palil
the valley of servoz
the cresta paganini
the hooker glacier
the adamello
the tiefenmatten
the festijoch
the brcithorn
the val di vitelli
the hochthaligrat

st. elias
the col gran neiron
the col de luisettes
the glacier de la breche
the val camonica
the col du loup
nantbourant
monte amaro
the valley of leuk
the mutthorn
the valsorey glacier
the tasch valley
col du chateau des dames
monte fernazza
the bachistock
col des planards
the glacier de monei
sommet des bouies
the fellaria glacier
the mont enchastraye
col des diablons
the vale of hasli
the festi glacier
trill glacier
the rofenkar glacier
the valle perse
the cijorenove glacier
the baltschieder glacier
boscolungo
the spranje
cima dei gelas
the urbachthal
the glacier de getroz
val vajoletti
the col de girardin
the glacier de bassac
faulhorn
the col de collon
the col de pila

the valeiglie
tre becchi
the oetzthaler ferner
the morteratsch glacier
the hufi glacier
thaler ferner
the val prato
col du lion
the dolomitic
the yerva pass
the dent du geant
the rochers des fiz
punta del dragons
the valley of zermatt
col de nivolet
the refuge de falp
the coritenza
the shreckhorn
the tuiber pass
col de valloire
the ruinette
the drei schuster spitze
cima del pizzo
the bisi thal
the herbetet
the geschenen thal
thecca di brenta
gemshorn
the aiguille de talfefre
the primiero mountains
de palpe
the sulaenthal
vallouise
the bergsehrund
the haut de cry
the munster alp
greenhorn
the petits mulets
the col de vars

the vallasco
the vallone di cougne
the col de sonadon
the passo di cavento
the col de la za
the darrei glacier
roche du bond basse
the bossons glacier
the mora alp
val della neve
the col del carro
piz bellavista
the flucnthorn
the balmenhorn
the viescherjoch
the como bianco
silber sattel
the schwarzenegg
the faudery ridge
the trogen alp
the wildspitz
the defereggen thai
the laquinjoch
the tossenhorn
ostalpen
the piz ner
the hanig alp
the vallon de la mariande
val di rabbi
aiguille de polset
the val cairos
the col de yallante
the cima del bousson
col de la fenetre
the sasaello pass
glacier de viesch
the bider glacier
bosenlaui
the gemelli della roccia viva
balme
the val di mel
the col de goleon
the pointe de rosablanche
the zwischbergen pass
punta michelis
the boval
the becca di luseney
the piz sella
pointe de mary
the col des trois pointes
col de castelnau
the grand veymont
the maquignaz
col de luisettes
stachelberg
the alphubelhorn
the dreisehusterspitz
the mont de la gouille
the etschthal
pizzo venezia
the col du bonhomme
the pierre a berenger
vallante
the col des courtes
the satchori pass
mt. fallet
val di bin
the cogne chamois
the fraele pass
the col de rochefort
the fomo glacier
the col des masses
the sella pass
piz medel
vreilerhorn
the glacier de pabeille
the col des grandes murailles
the val de rhymes
the montanvertat
the col ditelleccio
the paradis
the primiero valley
the glacier des eivettes
colle di sibolet
the vallon claus
the allchste spitze
the col della piccola
bachalp
the madatsch ridge
the glacier des ecandies
val rendena
the col de pl
piz bernina
the unterbachhorn
the schreck-horn
val popena alta
mont perdu
festi glacier
the glacier de girard
the glacier de freboutzie
the reichenbach
the gredetsch glacier
the etangons glacier
the glacier de parste
the eringerthal
val malenco
the za de zan glacier
the jungthal
colle traversette
the col de valsorey
the gaulis plateau
val di scalve
the bricolla glacier
the piz buin
the col de jaman
the valserhorn
the mittelhorn
the mont de la sengla
gamshorn
the valontey
the val di gesso
mont colon
thehorn
the crusehetta pass
the croix de belledonne
the glacier du brouillard
the thaileit spitz
the argentiere glacier
val nambino
the campitello
the cimes blanches
the sefinenthal
val federia
val lavinuoz
talefre
the col de sageroux
gwynant
the pic olan
the col de bellevue
rothhorn
col de la croix
the jilkisu
the maderaner
pointe du pisset
the vallon des salettea
val de lys
the val di valasco
piz quatervals
the pic de montandayne
val bendena
the chardon glacier
the antabbia glacier
the gassenried
the vieschergrat pass
the jagi glacier
the mont pilate

the monte di zocca
the renferhorn
the col de mary
the punta fiorito
the vallon de sellavieille
the basses alpes
the val de hibou
the hohes licht
the scerscen glacier
the col de creton
the rion glacier
the col del merlo
the vallais
the col de vallante
the glacier du grand
the cima di vallon
the herbetet glacier
the st. elias
san nicolo
the gletscher alp
col du thabor
monte rotta
the pic buille
the staffel alp
the pic de la grave
the trepalle
the maderaner thal
collon
the medel glacier
valle maggia
the foppa alp
the stafel alp
the piz tschierva
the silvretta glacier
the col du fond
the lauzon glacier
the verpeiljoch
the sehalliberg
the luschariberg
the breche glacier
the peuteret ridge
oberalpstock
col de la muande
the col di lago
the pena de oroel
the col de la sauce
the col della neve
colle del carro
the lutschine
monte bignone
val antigorio
the clos de la cavalle
val seriana
the vuibez glacier
piz roseo
the bossons
val noana
the ortler
the sella joch
the hennesiegel kopfe
piz buin
the weingarten glacier
the col de champorcher
the dreieckhorn
becca di nona
the col de sassa
the buffalora pass
the colline pisane
the theodul pass
the ausser locker spitze
the wengera alp
the scheinige platte
val maisama
the fassa thai
the graustock
the trou de toro
geierstein
the hornli ridge
the joch pass
the passo di ball
the sonadon
the col de lauteraar
the noaschetta glacier
the ulrichshom
the col bardoney
dent parassee
val di genova
val mergoscia
the salzkammergut
the glockthurm
the piz terri
the mur de la cote
the kammlistock
the aiguille de
the colle di merciera
mont mallets
col dee
val joufirey
the alp zarmine
the col de la pointe de bricolla
the glacier des etancons
the chamois pass
the passo rovano
the happen glacier
the roseg thal
the val champagny
the schwalbenkofljoch
val maygia
hanging glacier
the aiguille de peteret
the val di vallante
col de rioufroid
teniarhorn
the col de la gailetta
the st. gotthard
the gorner grat
the rocca del mat
val giuf
piz tavrii
the col de vauon pierre
the gelmerhomer
the verpeil spitze
the pic lory
grand col
the piz palu
the fresnay glacier
the pointe des ficrins
the la neuvaz glacier
the argentiere
the stockhorn
the colie dei piazzi
the plan de la cavalle
mont maudit
the cima bianca
the haas valley
the baifrinhorn
the col bonney
the val de la leisse
the parung pass
the geisspfad
monticella
the col de sais
the upper ara glacier
col ferrex
the basso di muro
the mont pelat
the punta di tersiva
the pujo glacier
the faschhorn
the balme des chamois
upper engadine
the konigsspitze
the grand colouret
the piz morteratsch
the gaulis
the civetta

the col de frette
the col des navettes
mont gruetta
the vogna
pichincha
the col de cian
the pointe des plines
the rosenlani glacier
monte gazza
the punta bianca
monte zovo
monte como
the val d'orco
the col de martignare
the petit coluret
allerhochste spitze
the ginevrie alp
the lanthorn
the glacier de la frasse
the pointe des salles
the col ferrex
col de la leisse
the col du palet
mont blanc
the glacier de l'epee
col becker
the dent du giant
the col de grandcrou
the glacier des ignez
the innthal
the col de belleface
the val di peccia
the puntota glacier
the rochers rouges
the ritterpass
the sasso rosso
the col de berard
la tour ronde
hornblende

the monch joch
the col de chalance
the becchi della tribulazione
val imagna
banhorn
piz kesch
dent de satarma
the pointe du colloney
the sonadon glacier
pointe de marguerite
the col de la brenva
the kulm
the val ciamoseretto
the flavona alp
the clemgia
the monte della disgrazta
the val de galambre
the col de pevsque
the grunhornllicke
the grat
the ober monch joch
the cima di piazzi
the lattenhorn
the becco di mezzodi
the collarin
the glacier de tabuchet
the valley of la berarde
the rienzjoch
the langthaler ferner
the glacier du casset
the schallen joch
the roth horn
monte spinale
the grivoletta
the glacier de miage
the glacier of ferpecle
the val mora
the lotschen lucke
bardoney glacier

the gelten glacier
the lotschsattel
monte blanc
monte tezio
the barenhorn
the monch
the gran neiron glacier
colle campaccio
the val anzasca
glacier de thomme
col emile pic
the altels
the turtmann range
the allerhochste spitze
the bagni di valdieri
the glacier de saleinaz
the tete du replat
the col de la vallette
the glacier des ignes
grand peak
mont roselette
the wellhorn
the ruchen pass
the punta budden
the bosco nero alp
the scale di fraele
the dent de corjon
the grands montets
the teleccio glacier
the montcnveis
the glacier de brenva
the engelhorner
the glacier des otanes
glacier de la grande
pena colorada
the aiguille du soreiller
pointe haute de mary
the forclaz
the maderanerthal

the yaleille glacier
upper brenva glacier
the mont gruetta
the plane de joux
nanzerthai
the col de fenstre
glacier de la casse deserte
the olden alp
col de cheville
the val masino
the grand coluret
the dent des bouquetins
the tete de la tribulation
the wengem scheideck
the kanderthal
the val d'orca
becca de ortton
fusshorn
val marson
the col de verbier
the valley
the alp di balme
hohenstollen
the alpe du tour
the mont joli
the arpettete
the val vermolera
the col du sele
the piz fliana
val antabbia
the bernetsmatt alp
colle della tribulazione
the glacier du grandcrou sud
pena de gorbea
the griineckhorn
the ebnefluh
the col de la casse deserte
the bruneggjoch
col de rochefort

the bonnepierre glacier
the val travignolo glacier
the tambohorn
the triftalp
monte vito
etzthaler ferner
the portienhorn
zilierthal
the grand pic
the valley of morgex
the madatsch neve
grand pic
the zocco pass
the punta forches
bonne pierre
the plattkofel
the piz languard
the gaderthal
the tour st. andre
the zmutt
the bertol glacier
the pic jocelme
val narcanello
vorspitze
the monte pian
mont clapier
the valasco
the corno bianco
the estellette
the fletsch joch
col de saleinaz
the valley of fee
the monte prese
col de galambre
mont herbetet
val jouffrey
the cima de jazi
the vrenelisgartli
the leukerbad
the einserkofel
pizzo stella
the nadel joch
the uschinen grat
the joch
the colle delle loccie
the zervreilerhom
val bajon
the kienthal
the broglio glacier
the nantulon glacier
the cresta agiuza
the col de garin
the piz urlaun
the glacier des etangons
val fassa
val lavaredo
col de la gasse deserte
bruneggjoch
the binnenthal
the tatelen plateau
the muretto pass
the brenta
fillar pass
the viescherhomer
the schilthorn
the vallon de la sausse
the hohberghorn
the tiefemnatten glacier
the ruchen
the vogel joch
the glacier di pujo
the punta giordani
the glacier de marinet
the col de la temple
the seisser alpe
the schwarz glacier
the borzago glacier
the scheerhom
the valley of queyras
the lawinen thor
col du grand sauvage
the silber sattel
the erstfeldthal
cima della rosetta
the hangend glacier
the cima della vezzana
the kerstelenbach
spitzbergen
the col mouei
monte civita
the dundengrat
the liiner joch
the teliecio glacier
the val de tinges
the col toumanche
elsigen glacier
the col des fours
birchfluh pass
the val brenta
the fiinffinger spitze
the sagis grat
the col de la cayolle
the aiguille blanche de peuteret
tchagerjochs
the col du tour
abricolla
piz roz
the telleccio glacier
the brenva glacier
monte foscagno
pic des agneaux
piz zupo
the punta foura
the val campaccio
the vallorgia glacier
the col de
val di fontane
the upper engadine
the chamonjx
the marjelen alp
the aiguille du glacier
the cogne valley
the bec du mont for
the val vecchia
col di val cournera
colouret
ochsenthal
the balmen-horn
the valsorey
the belalp
the pian del lenzuolo
bies joch
balmenhorn
the col du sellar
the val di forzo
the wengern scheideck
nant francon
the hochbalm glacier
the col de cheillon
the rocher de naye
the col du talon
olmenhorn
the grand pic de belledonne
the glacier d'invergnuon
the spondalunga
the oberalpstock
the glacier de zardesan
the val cavrein
the becca del merlo
the mont yentoux
the upper rosenlaui glacier
the colie di livoumea
the col de tignes
the montandevne glacier
ost spitze

the schreck horn
the becca di nona
the val del zebra
the doldenhorn
col la berarde
the val godemar
the spanorter joch
the vallone delle forciolline
the siidlenzspitze
the glacier du giant
the pointe de chamossaire
glacier de la bonne pierre
the sanetsch
vintschgau
the dorferkees
the col de la coste rouge
griinhorn
the vallon des
ortelerspitze
the sasso di ferro
konigsspitze
the pic tyndall
val travignolo
flavona alp
the pizzo tremoggia
the columbus glacier
the bemetsmatt alp
the gorner
col del porco
mont ruan
the corni del confine
gassijochf
the gauli glacier
the yintschgau
the brunni glacier
the vallon des etangons
the cima della nasta
the val de ribou
the charpoua glacier
the fee glacier
lower zardezan glacier
the sehalliberg glacier
the yiubez glacier
the col du tour noir
the grodener jochl
the pointe de la sana
the pierre berenger
monts de glace
the glacier de rochefort
the glacier du grand mean
the great horn
the glacier de charpoua
the schallhorn
the glacier de galese
the colle budden
the corno piccolo
the glacier of distel
the vallon de la pirade
the cima della culatta
val travernanzes
the monte maurigno
the aiguille du tacul
the bresciana alp
ponte di legno
piz linard
the col de giers
the colle di tenda
punta gastaldi
the nadeljoch
val di cruzzini
the fletschjoch
the pic d'otemma
the cima di brenta
the roccia viva glacier
the tuoi glacier
the cima viola
the ivan pass
the gliems glacier
furgen glacier
the valley of sixt
the monte croce
the fex glacier
the glacier du tabuchet
the vale of servoz
the bocca
the col de mont brute
col de chalance
the rhone valley
the grand tetret
the pointe percee
val di lei
veglia alp
the aiguille du gouter
the ciardonei glacier
val mastalone
the val del sasso
the pyrenees
the dlindengrat
sierra de cadi
the vallon de la muande
the col de la bren
the schaufelspitze
the valley of st. marcel
the alpe granus
the bee du grenier
the rossbodenhorn
col de la tour ronde
the val varaita
schweizer alpen
monte majella
the hohberg glacier
the val di rabbi
the col de pargentine
mont mounier
punta della uja
the pizzo
val pravitale primiero
the gepatsch glacier
the val bavona
the val di roda
the spitze
plate des agneaux
tour du grand st pierre
col de puissaille
the fiinffingerspitze
the val grisanche
valines
the val maisas
the aiguille blaitiere
the fellaria
bionnassa
the col de la scigne
the combe de savoie
val formin
the embors glacier
val buona
the petersgrat
the glacier de cheillon
the glacier de za de zau
the lammeren glacier
the paneveggio
the col de ni volet
val teresenga
breithorn
the chamois
the becca de nona
the bocca dei massddi
the lotschthal
the blumlis alp
the rostnlaui glacier
the pic du glacier
the walcherhorn
the oberhom alp
the mont chauve
the porchabella glacier
the lower gosau

the schmadri glacier
the spitzstein
the mesurina alp
the col de la portetta
the cevedale
the mestia pass
chste spitze
val orsine
the corbassiere glacier
the kali glacier
the val de cogne
the col de la casse
the oberaar joch
the gassenried glacier
montasio
the triest glacier
the aipitetta
the stauberbach
colle del becco
the combe de mary
the aiguille balmat
the ritter joch
the talefre
the valauria
the engadine
brunegghorn
sella pass
the val asinella
the pene blanche
the valnontey
the otemma glacier
the theodule
the punta giordano
the entremont
pic verdonne
thenenthal
the saleinaz glacier
the millaris plateau
val coumera

the nadelgrat
barrjoch
the ecandies
the mont fleuri
col des rouies
the arpisson
the leitergrat
the bocca di brenta
the val d'entremont
the dossenhorn
the pizzo valgrande
thalkirch
the diissistock
the col du lautaret
the col de baline
the pizzo del diavel
the schallihorn
the jamthal
the glacier de blaitisre
thecima di canale
the matterjoch
the ferpecle
the crete des bœufs rouges
the val tournanche
the simmelihorn
the glacier de moiry
the col de monciair
mount st-elias
the prievor ridge
the stelvio
the col de la lavey
the combe froide
val di non
the val sinestra
the valley of cairos
the grande plaque
the zumstein
the glacier du grand tetret
the brunegghorn

the samnaun valley
the tic du glacier
the schalliberg glacier
the trois freres
pointe de sainte anne
the boththal
miravalles
baltschieder
the val calanca
the ochsenthal
the sept laux
the upper aeschinen alp
the col de monei
the sandakphu
balme rousse
the oberalp see
roche moutonnee
the corbassiere
the schmadribach
the nuefelgiu pass
the como piccolo
the col delle sagnette
the cavergno glacier
the monte di castelli
the schwartzhorn
the grand glacier
the val delle seghe
monte gallina
the maya de bricolla
faderjoch
the mont falcon
the pic gaspard
the glacier de broglia
the corno del camoscio
the val lavinuoz
the glacier du geant
roche moutonne
the pointe de la yalettaj
the verra joch

the valley of llyn idwal
the upper alp
pic central
the aiguille de miage
monte di pietit
the fiitschol
the col de yalpelline
the kinsteraarhorn
di montasio
the cima del lago agnel
the moiry glacier
the aiguille
the pers glacier
the finsteraarjoch
monte cistella alta
monte di marte
monte nero
the meija
the liappey alp
the pisole alp
the fimberthal
the lower gail thal
the fum glacier
the combe
val codera
val camonica
the rochefort glacier
the tiefenmatten glacier
the val della rovina
the alpisella
the bergamesque mountains
the baltschieder joch
the furggenhorn
the val angrogna
the lautaret
tour noir
the monte oliveto
the biittlassen
the saleinaz

the cima dei gelas
the alp crozlina
doldenhorn
valsenestre
the pic bonvoism
the blumlis alp glacier
the col de jallorgues
the col de charnier
the colle di vanin
the allaleinhorn
the fenetre de dzasset
the hochbalen glacier
colico
the tinzenhorn
the montan
col des clochettes
the alpitie
the val della bovina
the col du vol
the faido
the val di mello
the hussein thal
the cappella di monte
val malvaglia
the col de mesoncles
the pointe de montandayne
val calanca
combe froide
val d'algone
the jungfraufira
sierra de gredos
the schwarzhorn
the val toumanche
col des cors
the vanescha pass
piz pisoc
the petit plateau
the alp gnof
the col de yay
the glacier de la tribulation
the col du celar
the criner furke pass
the col vert
the cima di lusiera
monte della stella
the bigerhorn
the allee blanche
glacier de salena
the val orsine
the val de lys
col de lauteraar
punta des cora
the yaleille
the col di s. bernardo
val sparlotsch
val d'arno
the palil glacier
the nantillons glacier
the col de champex
the croix blanche
the grand tour st. pierre
the basodino
the maloya pass
the grandes ascensions
the ulrichshorn
the hartley spitze
terhorn
piz yadret
the tiefenmatten joch
col vieux
the oberalp
the glacier carra
the urbach thal
the geislerspitzen
the triftgrat
the piz medel
the pala alp
the lampertsch alp
the vallon delle forciolline
val di avio
the glacier de gez
the russein thal
the pointe de sengies
monte cristallina
the turtman glacier
the pic verdonne
the piz bernina
the aiguille du gofitfi
the gabelhom glacier
the col de la lanze
the tambo glacier
the val maira
the glacier de la charpoua
the triftgletscher joch
the konig spitze
the aiguille de varens
the val champey
mont vinaigre
grasse combe
the angeluga alp
the tchagerjoch
the glacier de prou
the schweiben alp
val asinozza
the hautemma glacier
the fervall
aletsch pass
col de blancien
the alpe della bolla
the zardesan glacier
the punta vittoria
glacier of aletsch
the montagne de prou
the mont aliet
lower glacier
the punta del broglio
the zinnen
the care alto
the plachten alp
val masino
monte rosso
the fenetre
the saas thai
the col des ecandies
the ziilerthaler ferner
the cima
fellaria
the lysjoch
col di tenda
the punta di fontanella
the col de lauzon
the the brenva glacier
the fox glacier
the schailenjoch
the presena glacier
col perduto
the yentina glacier
piz borel
the cima di mercantoura
the fall glacier
the val de rhemes
the bezingi glacier
the scaletta glacier
the ampezzo
the tre coci pass
the pizzo san colombano
the grand peak
the rheinwald
the kistengrat
the llanberis
the gletscherhorn
the niederjoch
val mora
the viubez glacier
monte cavallo
the augstbord pass

the bietsch joch
the monte perdido
the hohberg pass
piz palii
the col du collarin
the ebihorn
the col vicentino
the val lavizzara
the upper verva
the walliser viescber glacier
the aroila glacier
val canzoi
the col de la grande luis
the gastereuthal
alpiglen
furggthal
the champatsch
the abberg glacier
the col de la gippiera
the glockhaus
the roches moutonnees
the col de la plate des agneaux
the portiengrat
the vallon
the col du says
the zigioronove glacier
the valley of castan
the campo di rutorto
mount st. elias
stellihorn
the alpe di veglia
the col de traversette
the blindenhorn
the hochste spitze
the mont de saxe
the col du lion
the col di finestre
the pre de bar glacier
the val di bosco

the barrjoch
the roseg valley
the col de mont rouge
the trient glacier
the col de la nouva
the montandeyne glacier
the misauna alp
vallon
the pointe de bricolla
the val sasso bisolo
the punta di ciampono
the piz eoseg
the col de eochefort
the cima de jazzi
the dreischuster
val belviso
the gabelhom
the pic de la pyramide
the presanella
the verstanklahorn
the cima della madonna
the val di brenta
the saasgrat
piz st. michael
the col de tracuit
the kuhe glacier
the glacier carre
the gspaltenhoraer
the abricolla
the findelthal
balferin glacier
the cima di nasta
the valle delle seghe
monte delle loccie
the saas-grat
the tabuchet glacier
the darboneire glacier
the wellenkuppe
the felik joch

the gomerer thal
the galenhorn
val di zoldo
the val lavaz
the glacier de lechaud
the sesiajoch
monte sissone
the col de galese
the col de rhymes
the glacier du fond
the cima del carro
the gasterenthal
the col dolent
col des aigles
val tournanche
the glacier de kothelsch
the ried glacier
the boval gite
leghorn
the chamonix valley
the montanvers
felik joch
the grenz glacier
the yalnontey
col de cretan
the col lombard
col du vallon
piz minger
the ochsenstock
the schreck glacier
monte folletto
the glacier des bassons
the caucasus
val bavona
the monte nero
the lenta glacier
val brenta
val chigniulascio
the cima di fradusta

the mont corve glacier
val cabione
the zimbaspitz
the vallon des bancs
the riflelhorn
alp della neve
val cluozza
scerscen pass
the glacier de cijordnove
mont cheillon
the col de fours
bieshorn
the nagler spitz
nant gwynant
fillar joch
the aiguille de la yola
the lauteret
the lotsch pass
linththal
the val jouffrey
the lys-joch
the tete blanche
col des pousses
the grande sagne
the hangendhorn
camonica
the beichgrat
the yuibez glacier
the funffingerspitze
monte leone
piz margna
the almagel alp
the mitteljoch
the val cantone di
col di telleccio
piece glacier
piz foraz
the furggthal
the grand tour

the yesilspitz
the tschingellochlihorn
monte pian
the einfischthal
vallette
the vin du glacier
unterbachhorn
the weitenalpstock
the val piantonello
the col de lautaret
the zumstein spitze
val pravitale
col de la coste rouge
the glacier d'orny
the mont de la brenva
the rouies
the balenfirnjoch
rinderhorn
the grande fourche
monte gristallo
valdieri
the valletta
the hexenkopf
the huddleston glacier
roche chevriere
the cristallo ridge
col de roche mantel
the astazou
the jeggischhorn
the val maggia
the venediger
the mellichen glacier
the ausser barrhorn
val kiva secca
pics gastaldi
the montandayne glacier
col du sirac
the piode glacier
the mont perdu

the fenetre de salena
val leventina
the glacier de l'arpont
the dungel glacier
the ost spitze
arpette glen
the val di boda
col des aravis
the arbe glacier
the vallon laugier
col brusau
the rosenlaui glacier
the lebendun
the val cengia
the cedeh glacier
the valley of trient
the moming pass
the est spitze
the monte della stella
the barenalp
thegeler alp
the piz mokteratsch
the col de la muande
the cima cadino
the montagnaia
the col des hirondelles
the val de st. marcel
the drei zinnen
the mont aiguille
the val malenco
the lotschenthal
the bietschthal
the tabaretta thal
the col du
val anziei
the buss alp
the col des serins
the zervreilerhorn
pieve di cadore

monte durano
the alp kobiei
the rothblatt glacier
the ban glacier
the upper brenva glacier
monte matto
montenvers
the zagen glacier
monte pisonet
the fedaja pass
egginerhorn
col des hirondelles
the val vraita
the dome du goute
trotzi pass
val di susa
the monte bignone
col des grandes bousses
the aiguille du plan
the brunni alp
the pic de retour
the clos des cavales
the glacier de tronchey
the furgge glacier
monte augusta
the val stura
the croda di bagion
the mantliser grat
monte oristallo
the aiguille du midi
hinterrhein
piz murtarol
the val tellina
the val di braulio
the langthaler joch
piz palu
the hasli thal
the breit-horn
the torre di brenta

the col du thabor
becca conge
the nantillon glacier
tour de creton
eggerofen arbola
the glacier des amethystes
the valley of sexten
the altels glacier
the ampezzo valley
the riffelberg
rizlihorn
val di mel
the col di telleccio
the cogne mountains
the crozlina alp
col de goleon
the rocher blanc
the ecandies glacier
glacieres
the rossboden
col verdonne
the pian del cavallo
the val carapiglia
the riffelhorn
the arbola glacier
the sulzenan glacier
the grindel alp
the matschacher
val campedelle
the valtelline
the konigs spitze
the aiguille de blaitiere
the lauteraarjoch
the telleccio
valley of grisanche
monte caggio
matterhorn
pointe du mulinet
the glacier de sengie

monte eosa
the vallonpierre
the grosse alp
punta sengie
monte ventina
pena de oroel
the alp
the col du charforon
the tschierva glacier
the bashil
the verra glacier
the petit mont colon
the schwarzberg
the valley of lanzo
valloire
the ampezzo dolomites
the scalare di telleccio
the piz scharboden
the valley of ice
the glerscherhorn
the val fiera
the madatsch joch
the corno di dosde
the dossradond pass
the mont des agneaux
the var valley
val di prato
mont tinibras
the monei glacier
mont emilius
piz formin
the galenstock
val di lucerna
the col de joux
the col di tenda
val ombretta
glacier des ecrins
the val devero
the cima di jazzi
the noasehetta glacier
the punta sella
the bies glacier
glacier du clot des cavales
the mont lellard
the saas valley
the gasteren thal
the pont de mauvoisin
the glacier des roches
val fonda
the clocher du lac
val vermiglio
the col du midi
the mischabelhorner
the rofel pass
valtellina
the wischberg
the pointe de garin
the matscher thai
the caputschin pass
mont pelat
monte rosa
the pizzo campo tenca
the wolfsbachthal
sckontauf spitze
sierra de bejar
col du petit
the val lazin
augstbord pass
the lotsch glacier
barrhorner
the col des ecrins
the combeynot
the glacier de zardezan
the bee de phomme
the glacier du bouchet
the val de viu
the col de bardoney
the julier pass
monte forato
val challant
val ruvinian
the sageroux
the val pradidali
the col de la croix
the pointe de tinneverges
the yispthal
the zmutt ridge
the bies joch
the balferin glacier

Annexe 4
Intersection des données entre les résultats de DistilBERT et de spaCy

the bel-alp
the col de la neuvaz
val toumanche
the grohmann glacier
the val de fournel
the zupo pass
the vallon de tinibras
the col de la galise
the vigne alp
the combe di valleiglia
col de claire
the glacier de durand
the mont thabor
the val gordolasca
the vallone del roc
the wajtersfirren alp
the pic central
val di lanzo
the ried pass
val formazza
the col de la lune
the monte bosa
the cima del caire cabret
the glacier de bertol
the s. madatsch spitze
the allelein pass
the alpien glacier
col des rousses
val canali
the erstfeld thal
the col du petit gl
the alp di ferro
the col des cavales
the pointe du pousset
la balme
the col du piolet
the trelatete glacier
the val niiglia
the bricolla alp
the lebendun pass
val cengia
zervreilerhom
the pic du midi
the aiguille de berenger
the montagne des coeurs
the grandcrou glacier
the glarnisch
the gallo pass
vin du glacier
the alpe ravina
the val de peychouda
monte kosa
the val giuf
the col de la breya
the aiguille du plat
val anzasca
the col de raus
the val tuoi
the fenetre de cogne
the hied glacier
col budden
the col de fenetre
val del sasso
piz vadred
the pointe de mary
the col du tacul
the glacier de lauzon
the wingern alp
piz vadret
val grisanche
the glacier de chauvet
the otro valley
val verzasca
the pointe de marguerite
the col de vallon pierre
piz morteratsch
col du cormet
the pointe du bousson
monte zigolo
the vallon de chillol
the kistengrat pass
the drakosh pass
the pas de cheville
the rhone glacier
serravalle
piz chalchagn
the col des bouquetins
the monte eosa
the scaletta pass
the cristallo glacier
the alpisella pass
viescherhorn
becca di monciair
aiguille du midi
the stein alpe
the biferten glacier
grohmann spitze
mont albert
the col de blancien
the val fassa
mittelhorn
the val de la gitte
the schwarze glacier
the val duron
glacier de merddre
val lavizzara
pied du col
schallen joch
the glacier de la bonne pierre
the monciair glacier
the glacier of valleiglia
the val peisey
col de la lune
the amethystes glacier
the monte plessura
the glacier de tacconai
the aiguille de la peau blanche
gebelhorn
the glacier de nant blanc
valaisan
the col de mont brule
the bresciana glacier
gassispitz
the val des navettes
the pointe de ceresole
val fiera
val vigornesso
the col d'ormelune
val di forzo
the padeon alp
the glacier de yaudet
the col de planereuse
the jung pass
la grande serre
rhone glacier
val ostera
the netz glacier
the farno glacier
stellihorn
the col de galdse
the glacier de bouquetins
the aiguille de charmoz
the grodner joch
pic jocelme
the col de girarain
the dents des bouquetins

val devero
colle budden
the zebles pass
blinnenhorn
the col de la pilatte
monte vibo
the col de brouis
the val des bans
the piz humor
piz plavna
the glacier de grandcrou
the hohsand glacier
the vincent pyramid
yispthal
the col di cerieja
the dolomites
the glacier de monestier
the tasch alp
the fillar alp
the pic de neige cordier
the gross ruchi
trippachthal
the grand glacier bellaza
mont rose
the salena glacier
vispthal
the pioda alp
the pas de turloz
dents des bouquetins
the val zebra
mont glapier
fluchthorn
colle fiorito
val maggia
the moncorve glacier
mont joli
piz ltnard
mount ararat
the val de grauson

the alp oberkasern
the glacier de fos
col de la lavet
col des trois pointes
the becca di noaschetta
the glacier de grandcrou sud
the bernina range
the col du lauteret
the pic du frene
the col de la maigna
the tour noir glacier
the pointe de yaleille
the glacier de pierre joseph
gassispitzen
the val ferret
the col de nivollet
the tete noire
vallombrosa
the wengem alp
the inner lockerspitz
the estellette glacier
col de valsorey
rhein thal
turtmanthal
val livigno
the roccia viva
the zapport alp
the grand yeymont
the petit mont
val yerzasca
montets
the valle grande
the turtmann glacier
chermontane
the gietroz alp
the val verzasca
the siidlenz spitz
the col du couard
valetta

the col du gdant
mutthorn
the val di canale
the col des aravis
dolomites
col de fenstre
the pic de bure
aiguille de la sausse
the glacier du mulinet
the pizzo columbe
the col cliamonin
tete noire
the mont colon
the val campiglia
the grosse windgelle
the binnen thai
saasthal
the vertain spitze
schallhorn
the val des ormonds
the aiguilles rouges
the piz linard
mont lacha
the piz formin
the col theodule
the valle de viil
monte bocchetta
val grana
valsorey
the pic oriental
monte bego
monte mottarone
the col de la croix haute
the pic du thabor
the allalin glacier
glacier de fenetre
monte morrone
col de la sauce
the val fomo

rossbodenhorn
the cima di canale
rocher rouge
the saas thal
the col di teleccio
the nagles spitze
the glacier de la casse deserte
the schrotter joch
the monte di scerscen
the pizzo di ferro
the blindseeli alp
the tour st. pierre
the sissone glacier
the monte rotta
the bocca dei camuzzi
the piz zupo
monte eotta
the glacier des nantillons
glacier de freboutzie
the val de pagnelle
the col vaudet
the glacier de chaviere
the col de luseney
nantillon
the val del leno
col du gyant
the val sorapis
the glacier de vaudet
the val travignolo
the weiten alp
mont de lans
mont brule
pizzo campo tenca
the richetli pass
the plattas alp
the glacier de gai
the val scura
hangendhorn
the val grosina

mont-blanc
the glacier of ayas
the yal mora
the piz palil
the hooker glacier
the val di vitelli
valtelline
st. elias
the col de luisettes
the glacier de la breche
the val camonica
the col du loup
monte amaro
col du chateau des dames
monte fernazza
the glacier de monei
the fellaria glacier
col des diablons
the festi glacier
col gran neiron
the valle perse
the baltschieder glacier
cima dei gelas
val vajoletti
the glacier de getroz
the eien alp
the col de girardin
the glacier de bassac
faulhorn
the col de collon
pointe de mandalon
the col de pila
the oetzthaler ferner
the morteratsch glacier
the hufi glacier
the val prato
col du lion
the yerva pass
col de nivolet
the tuiber pass
col de valloire
the drei schuster spitze
cima del pizzo
the bisi thal
the herbetet
the geschenen thal
gemshorn
vallouise
the munster alp
the col de vars
the col de sonadon
the col de la za
the darrei glacier
the sanfleuron glacier
the bossons glacier
liappey alp
the col del carro
piz bellavista
the st. giacomo pass
saas thai
the como bianco
the trogen alp
the defereggen thai
the pic de marbora
the piz ner
the hanig alp
the vallon de la mariande
val di rabbi
aiguille de polset
yesulspitz
the col agnel
the val cairos
the col de yallante
col de fenetre
col de la fenetre
the sasaello pass
the col de goleon
the pointe de rosablanche
the zwischbergen pass
the becca di luseney
the piz sella
pointe de mary
col de castelnau
pizzo venezia
the col du bonhomme
vallante
the col des courtes
the col de rochefort
the fraele pass
the fomo glacier
the col des masses
the sella pass
the glacier de gietroz
piz medel
the glacier de pabeille
the col des grandes murailles
the val de rhymes
the glacier des eivettes
the vallon claus
the col della piccola
val rendena
piz bernina
the glacier de freboutzie
the gredetsch glacier
the glacier de parste
val malenco
val ferret
rothhom
the col de valsorey
val di scalve
the col de chermontane
the piz buin
the col de jaman
the val di gesso
col de la roche tfalvau
the val agnola glacier
mont colon
the crusehetta pass
the croix de belledonne
the glacier du brouillard
glacier du triangle
the thaileit spitz
the argentiere glacier
val nambino
the cimes blanches
hintereis glacier
val federia
val lavinuoz
the col de sageroux
the cols longet
the pic olan
the col de bellevue
col de la croix
the maderaner
the piz casana
val de lys
the val di valasco
piz quatervals
the pic de montandayne
val bendena
the bee de mont forchu
the chardon glacier
the antabbia glacier
the monte di zocca
the col de mary
the vallon de sellavieille
the planail thai
the val de hibou
the scerscen glacier
the col de creton
the col del merlo
the col de vallante
the glacier du grand
the herbetet glacier
the st. elias
the gletscher alp

col du thabor
monte rotta
the staffel alp
the pic de la grave
the maderaner thal
the medel glacier
the foppa alp
the piz tschierva
the silvretta glacier
the col du fond
the lauzon glacier
the breche glacier
the peuteret ridge
col de la muande
the col di lago
the col de la sauce
monte bignone
val antigorio
the clos de la cavalle
val seriana
the vuibez glacier
piz roseo
val noana
the sella joch
piz buin
the weingarten glacier
the col de champorcher
becca di nona
the col de trelatetc
the col de sassa
the buffalora pass
the ausser locker spitze
the wengera alp
val maisama
the fassa thai
the trou de toro
the joch pass
the col de lauteraar
the stalden joch
the noaschetta glacier
the col bardoney
dent parassee
val mergoscia
the piz terri
the col du grand tetret
the essettes ridge
the mur de la cote
val joufirey
the alp zarmine
the col de la pointe de bricolla
the chamois pass
the passo rovano
the happen glacier
the roseg thal
the val champagny
the embours thal
the aiguille de peteret
the val di vallante
col de rioufroid
the col de la gailetta
the gorner grat
val giuf
piz tavrii
the col de vauon pierre
the verpeil spitze
the pic lory
the fresnay glacier
the la neuvaz glacier
the plan de la cavalle
mont maudit
the cima bianca
the col bonney
the val de la leisse
the parung pass
the col de sais
the basso di muro
the mont pelat
the pujo glacier
val delle seghe
the grand colouret
the piz morteratsch
the col de frette
the pointe des plines
monte gazza
the punta bianca
monte zovo
monte como
the col de martignare
the petit coluret
allerhochste spitze
the ginevrie alp
the glacier de la frasse
the pointe des salles
the col ferrex
col de la leisse
the col du palet
col becker
the col de grandcrou
the glacier des ignez
the col de belleface
the rochers rouges
the col de berard
la tour ronde
the monch joch
the col de chalance
the val marson
the becchi della tribulazione
val imagna
banhorn
piz kesch
the pointe du colloney
the sonadon glacier
pointe de marguerite
the col de la brenva
the glacier du yallon
the val ciamoseretto
the flavona alp
the glacier de trelatete
the val de galambre
the col de pevsque
the becco di mezzodi
the glacier de tabuchet
the bliimlis alp
the langthaler ferner
the glacier du casset
monte tublan
monte spinale
the glacier de miage
the val mora
bardoney glacier
the gran neiron glacier
monte blanc
colle campaccio
col des essettes
piz michel
glacier de thomme
col emile pic
the allerhochste spitze
the glacier de saleinaz
the glacier des ignes
mont roselette
the bosco nero alp
the pic du marbore
the dent de corjon
the teleccio glacier
the ofen pass
the glacier de brenva
glacier de la grande
the aiguille du soreiller
the yaleille glacier
the col de fenstre
glacier de la casse deserte
the olden alp
the val masino
les ecrins
the grand coluret

the haute balagne
becca de ortton
val marson
the col de verbier
the alp di balme
the yitelli glacier
the val de champoteon
the alpe du tour
the val vermolera
the piz fliana
val antabbia
the bernetsmatt alp
the glacier du grandcrou sud
the col de la casse deserte
col de rochefort
monte vito
the tierser thai
etzthaler ferner
the grand pic
the zocco pass
the piz languard
the tour st. andre
monte popena
the bertol glacier
the pic jocelme
val narcanello
the monte pian
mont clapier
the corno bianco
spannorter joch
the monte prese
col de galambre
mont herbetet
the tour grand st. pierre
val jouffrey
the cima de jazi
pizzo stella
the uschinen grat
val bajon

the broglio glacier
the nantulon glacier
dreiliinderspitz
the cresta agiuza
the col de garin
the piz urlaun
the val viola
val fassa
val lavaredo
bruneggjoch
the blinnen glacier
the grandes dents de veisevi
fillar pass
the albrun pass
the vallon de la sausse
the vogel joch
the punta giordani
the glacier de marinet
the col de la temple
the seisser alpe
the schwarz glacier
the col de galambre
the borzago glacier
col du grand sauvage
mont capucin
the hangend glacier
the aiguille du pcan
bondaaca glacier
spitzbergen
monte civita
col de chermontane
glacier de grancrou
the col des fours
the val de tinges
the col toumanche
birchfluh pass
the fiinffinger spitze
the aiguille blanche de peuteret
the forzo valley

the col du tour
piz roz
the telleccio glacier
the brenva glacier
monte foscagno
pic des agneaux
piz zupo
the punta foura
the vaieiglia glacier
the val campaccio
val di fontane
the stallenkopf pass
the aiguille du glacier
the val vecchia
ochsenthal
the gamchi glacier
bies joch
balmenhorn
the col du sellar
the val di forzo
the wengern scheideck
the hochbalm glacier
the col de cheillon
the rocher de naye
the col du talon
the glacier d'invergnuon
the glacier de zardesan
the val cavrein
the col de tignes
the ritter alp
the montandevne glacier
ost spitze
glacier de la tombe murec
the becca di nona
the val del zebra
vallesina
the glacier du giant
the pointe de chamossaire
glacier de la bonne pierre

the col de la coste rouge
the col collon
the pic tyndall
val travignolo
flavona alp
the pizzo tremoggia
the columbus glacier
the bemetsmatt alp
the col de zarmine
mont ruan
the gauli glacier
the col de lion
the brunni glacier
the val de ribou
the charpoua glacier
the sehalliberg glacier
the col du tour noir
the pointe de la sana
the glacier de rochefort
the glacier du grand mean
the glacier de charpoua
the colle budden
the glacier de galese
the corno piccolo
val travernanzes
the monte maurigno
the bresciana alp
piz linard
the col de giers
the pic d'otemma
the roccia viva glacier
the tuoi glacier
the cima viola
the gliems glacier
the col du conard
furgen glacier
the glacier du tabuchet
the col de mont brute
col de chalance

the piz umbrail
the rhone valley
the pointe percee
the glacier of salena
veglia alp
the aiguille du gouter
the ciardonei glacier
the val del sasso
the alpe granus
the bee du grenier
the val varaita
the hohberg glacier
the val di rabbi
the col de pargentine
mont mounier
the gepatsch glacier
the val bavona
valsavaranche
the val di roda
col de puissaille
the val grisanche
the val maisas
the col di monei
la combe
the col de la scigne
val formin
val buona
the glacier de cheillon
the glacier de za de zau
the glacier de la grande serre
val teresenga
monch jochon
the becca de nona
the blumlis alp
the pic du glacier
the oberhom alp
the vertrain spitze
the porchabella glacier
the schmadri glacier
the mesurina alp
the col de la portetta
the mestia pass
val orsine
the corbassiere glacier
the val de cogne
the col de la casse
the dufour spitze
the aiguille balmat
brunegghorn
sella pass
the otemma glacier
pic verdonne
the saleinaz glacier
val coumera
the pitz thai
the mont fleuri
col des rouies
the val d'entremont
the pizzo valgrande
mahrenhorn
the col du lautaret
the col de baline
the pizzo del diavel
the glacier de blaitisre
the val tournanche
the col torrent
the glacier de moiry
the col de monciair
the tour de grauson
the mauvais pas
valnontey
the col de la lavey
the combe froide
the val sinestra
the grande plaque
the mont maudit
the glacier du grand tetret
the tic du glacier
the schalliberg glacier
pointe de sainte anne
the val calanca
the col de monei
col des navettes
mont collon
the oberalp see
the nuefelgiu pass
the como piccolo
the cavergno glacier
the monts maudits
the grand glacier
monte gallina
the maya de bricolla
the vallon de lanchatra
faderjoch
the mont falcon
the pic gaspard
the glacier de broglia
the corno del camoscio
the val lavinuoz
the glacier du geant
the pointe de la yalettaj
pic central
the aiguille de miage
monte di pietit
the col de yalpelline
the cima del lago agnel
the pers glacier
monte cistella alta
monte nero
the liappey alp
the pisole alp
the fum glacier
the aiguille blanche de
val codera
the gros crenier
val camonica
the rochefort glacier
the tiefenmatten glacier
the baltschieder joch
the val angrogna
tour noir
the monte oliveto
the parrot-spitze
the alp crozlina
doldenhorn
valsenestre
the pic bonvoism
the blumlis alp glacier
the col de jallorgues
the col de charnier
the hochbalen glacier
vallauris
the col du vol
the val di mello
col des bouquetins
the hussein thal
the cappella di monte
val malvaglia
the col de mesoncles
the pointe de montandayne
val calanca
val d'algone
the val toumanche
piz pisoc
the petit plateau
the alp gnof
the glacier de la tribulation
the col du celar
the criner furke pass
the col vert
the val orsine
the val de lys
col de lauteraar
val sparlotsch
mount hell
rochers rouges

the palil glacier
the nantillons glacier
the col de champex
the croix blanche
the eringer thai
the tete de la maye
the maloya pass
the hartley spitze
piz yadret
col vieux
the glacier carra
the urbach thal
the piz medel
the lampertsch alp
the russein thal
the pointe de sengies
monte cristallina
the turtman glacier
the pic verdonne
the piz bernina
the aiguille du gofitfi
the gabelhom glacier
the col de la lanze
the val maira
the glacier de la charpoua
val tuors
the konig spitze
the val champey
mont vinaigre
the glacier de prou
the schweiben alp
val asinozza
aletsch pass
col de blancien
the zardesan glacier
lower glacier
the punta del broglio
the plachten alp
val masino
monte rosso
the col des ecandies
the ziilerthaler ferner
the col de lauzon
the fox glacier
the presena glacier
the yentina glacier
the po valley
piz borel
the fall glacier
the bezingi glacier
the scaletta glacier
the tre coci pass
the pizzo san colombano
gras alp
val mora
the viubez glacier
monte cavallo
the augstbord pass
the bietsch joch
the monte perdido
col burne
the hohberg pass
piz palii
the col vicentino
the val lavizzara
the aroila glacier
val canzoi
the col de la grande luis
furggthal
the piz spigna
the abberg glacier
the col de la gippiera
the col de la plate des agneaux
the gross doldenhorn
the col du says
piz ault
mount st. elias
the ddme du gofiter
the col de traversette
the hochste spitze
the mont de saxe
the col du lion
the col di finestre
alphubelhorn
vallesinella
the pre de bar glacier
the great tower
the col de mont rouge
the col de la nouva
the col jaillet
the misauna alp
the montandeyne glacier
the pointe de bricolla
mont aiguille
the piz eoseg
the glacier de m. durand
the col de eochefort
the cima de jazzi
val belviso
piz st. michael
the col de tracuit
the glacier carre
the puy de ddme
pointe des plines
the pointe des papillons
the felik joch
the gomerer thal
the s. ridge
val di zoldo
the val lavaz
the glacier de lechaud
monte sissone
the col de galese
the col de rhymes
the glacier du fond
the col dolent
val tournanche
the herbetet s. ridge
the ried glacier
leghorn
felik joch
the grenz glacier
col de cretan
the col lombard
col du vallon
piz minger
the pic des areas
monte folletto
val bavona
the monte nero
the lenta glacier
val brenta
col des ecrins
val chigniulascio
monte pizzoc
val cabione
the vallon des bancs
val cluozza
the col de la fenetre
scerscen pass
the glacier de cijordnove
mont cheillon
bieshorn
the nagler spitz
fillar joch
the aiguille de la yola
linththal
the val jouffrey
the tete blanche
the petite pointe de planereuse
the yuibez glacier
monte leone
the almagel alp
the vai savaranche
the grand tour
monte pian

vallette
the vin du glacier
the fee alp
the val piantonello
the col de lautaret
the zumstein spitze
val pravitale
rinderhorn
the grande fourche
monte gristallo
the huddleston glacier
the val maggia
the montandayne glacier
val de grauson
col du sirac
the mont perdu
val leventina
the glacier de l'arpont
the dungel glacier
the ost spitze
arpette glen
the val di boda
col des aravis
the la neuvax glacier
the rosenlaui glacier
the cedeh glacier
the moming pass
the est spitze
the col de la muande
the cima cadino
the col des hirondelles
the val de st. marcel
the drei zinnen
the val malenco
the tabaretta thal
val anziei
the buss alp
the mont tabel glacier
monte durano
the alp kobiei
the ban glacier
the upper brenva glacier
monte matto
montenvers
the zagen glacier
the fedaja pass
egginerhorn
the val vraita
val di susa
the monte bignone
col des grandes bousses
the valeille glacier
the aiguille du plan
the brunni alp
the pic de retour
the clos des cavales
the glacier de tronchey
the col du g£ant
the furgge glacier
monte augusta
the val stura
tuoi glacier
monte oristallo
val tuoi
the aiguille du midi
the scatta minojo
piz murtarol
the val tellina
the val di braulio
the langthaler joch
piz palu
the hasli thal
the col du thabor
becca conge
the nantillon glacier
the glacier des amethystes
the ampezzo valley
val di mel
rizlihorn
the col di telleccio
the cogne mountains
the crozlina alp
steinthalhorti
the gross spanort
the rocher blanc
col verdonne
the val carapiglia
the arbola glacier
the sulzenan glacier
the grindel alp
val campedelle
the konigs spitze
the aiguille de blaitiere
val tellina
the lower glacier
monte caggio
the glacier de sengie
monte eosa
the tour d'arpisson
monte ventina
the col du charforon
the tschierva glacier
the verra glacier
the petit mont colon
the ampezzo dolomites
the val fiera
the yal malvaglia
the dossradond pass
the corno di dosde
val di prato
the monei glacier
mont emilius
piz formin
val di lucerna
the col de joux
the col di tenda
the glacier des doves blanches
val ombretta
glacier des ecrins
the val devero
the noasehetta glacier
the bies glacier
glacier du clot des cavales
the gasteren thal
the pont de mauvoisin
val fonda
val vermiglio
the col du midi
the col du g6ant
valtellina
the pointe de garin
the matscher thai
mont pelat
the pizzo campo tenca
the val camperi
sckontauf spitze
the col de dza
the val lazin
augstbord pass
the glacier de zardezan
the bee de phomme
the glacier du bouchet
the val de viu
the col de bardoney
monte forato
val challant
val ruvinian
the val pradidali
the pointe de tinneverges
the zmutt ridge
the bies joch

Annexe 5
Intersection des données entre les résultats de BERT et de spaCy

the col de la neuvaz
the grohmann glacier
the val de fournel
the zupo pass
the col de la galise
eschinen alp
the vigne alp
gorner
the combe di valleiglia
col de claire
the glacier de durand
the aiguille de bionassay
the mont thabor
the mont rouge
the val gordolasca
the vallone del roc
the val d'aussois
the wajtersfirren alp
the pic central
val di lanzo
the ried pass
val formazza
the col de la lune
the monte bosa
the cima del caire cabret
the val formazza
the glacier de bertol
engelhorner
the allelein pass
the val di mont clapier
col des rousses
val canali
the erstfeld thal
the alp di ferro
the col des cavales
the pointe du pousset
becca di luseney
la balme
the grasstaller ferner
great scheideck
the col du piolet
the anderegg joch
the val niiglia
gasthaus piz ureza
the bricolla alp
the lebendun pass
val cengia
zervreilerhom
the pic du midi
the aiguille de berenger
the montagne des coeurs
the grandcrou glacier
the glarnisch
the gallo pass
vin du glacier
the upper fluchthorn glacier
the alpe ravina
the val de peychouda
monte kosa
arpette
the val giuf
the fellaria alp
the col de la breya
the aiguille du plat
val anzasca
the col de raus
the val tuoi
the fenetre de cogne
the hied glacier
col budden
the col de fenetre
val del sasso
piz vadred
the pointe de mary
the col du tacul
the glacier de lauzon
the wingern alp
the planail ferner
piz vadret
val grisanche
the glacier de chauvet
val verzasca
the becca di mezzo di
the col de vallon pierre
piz morteratsch
col du cormet
col de tracuit
monte zigolo
the dents de bertol
the vallon de chillol
the great scheideck
the kistengrat pass
the drakosh pass
saleinaz glacier
the rhone glacier
the tour st ours
serravalle
piz chalchagn
the monte eosa
the scaletta pass
the cristallo glacier
the tete du rouget
viescherhorn
the groden valley
the glacier lombard
becca di monciair
the glacier de bonnepierre
the stein alpe
aiguille du midi
the biferten glacier
grohmann spitze
the piano del re
mont albert
breche de la charriere
the col de blancien
mittelhorn
the val fassa
the val de la gitte
the schwarze glacier
colle maurigno
val lavizzara
pied du col
the glacier de la bonne pierre
tete de valnontey
the monciair glacier
the val peisey
col de la lune
the amethystes glacier
the monte plessura
glacier de chauvet
the glacier de tacconai
the aiguille de la peau blanche
gebelhorn
the glacier de nant blanc
valaisan
the col de mont brule
the bresciana glacier
gassispitz
the col de pinfemet
val fiera
val vigornesso
val di forzo
the padeon alp
monte montagnaia
the glacier de yaudet
the ciamoseretto glacier

piz lischanna
the jung pass
rhone glacier
val ostera
the netz glacier
the farno glacier
stellihorn
the col de galdse
the col de girarain
the aiguille de charmoz
the grodner joch
pic jocelme
colle budden
the dents des bouquetins
val devero
the zebles pass
the col de la pilatte
monte vibo
the col de brouis
the val des bans
the piz humor
piz plavna
the glacier de grandcrou
the hohsand glacier
the vincent pyramid
the col di ranzola
yispthal
the col di cerieja
the dolomites
the glacier de monestier
the marzoll ferner
the tasch alp
the fillar alp
the pic de neige cordier
the gross ruchi
mont rose
the glacier del broglio
the salena glacier
vispthal
the pioda alp
the pas de turloz
the val zebra
fluchthorn
colle fiorito
val maggia
the perdu glacier
the moncorve glacier
mont joli
mount ararat
the mountains of cogne
the val de grauson
the glacier du nant blanc
the alp oberkasern
the glacier des lancettes
the glacier de fos
the sass del eepos
col des trois pointes
the becca di noaschetta
the glacier de grandcrou sud
aiguilles de luisettes
the col du lauteret
aiguilles rouges
the pic du frene
the col de la maigna
the pointe de yaleille
the glacier de pierre joseph
the pizzo pobcellizzo
gassispitzen
the val ferret
the col de nivollet
the tete noire
vallombrosa
the wengem alp
tre croci
the estellette glacier
the aiguille de tronchey
col de valsorey
the breslauer hiitte
rhein thal
val livigno
the roccia viva
the zapport alp
col du vallon laugier
the petit mont
val yerzasca
montets
aigues rousses
the valle grande
the turtmann glacier
the gietroz alp
the glacier de toule
the siidlenz spitz
the col du couard
valetta
the col du gdant
mutthorn
the val di canale
the aiguille de pepaisseur
the col des aravis
dolomites
col de fenstre
the pic de bure
the sellinen tobel
aiguille de la sausse
the glacier du mulinet
the pizzo columbe
the col cliamonin
the stein glacier
tete noire
the mont colon
the val campiglia
the grosse windgelle
saasthal
the vertain spitze
schallhorn
the val des ormonds
the piz linard
mont lacha
the piz formin
the col theodule
monte bocchetta
val grana
the vallon de beauvoisin
valsorey
the pic oriental
monte bego
monte mottarone
cima verdona
the col de la croix haute
the pic du thabor
the allalin glacier
monte morrone
col de la sauce
the val fomo
the alp gulma
the cima di canale
rocher rouge
the saas thal
the col di teleccio
the nagles spitze
the glacier de la casse deserte
the schrotter joch
the cunonega alp
the pizzo di ferro
the tour st. pierre
the sissone glacier
montandeyne
the monte rotta
the bashilsu glacier
the bocca dei camuzzi
the piz zupo
monte eotta
the glacier des nantillons
glacier de freboutzie
the val de pagnelle
the col vaudet

the glacier de chaviere
the russein alp
the col de luseney
the val del leno
the val sorapis
the glacier de vaudet
the val travignolo
the weiten alp
mont de lans
the mutsch glacier
mont brule
pizzo campo tenca
the richetli pass
the plattas alp
the glacier de gai
the val scura
hangendhorn
the val grosina
mont-blanc
the glacier of ayas
casale di san nicolo
the cercen pass
the piz palil
forno pass
the valles pass
the hooker glacier
the val di vitelli
st. elias
glacier superieur des agneaux
the col de luisettes
the aiguille de peuteret
the glacier de la breche
the val camonica
the col du loup
monte amaro
col du chateau des dames
monte fernazza
the glacier de monei
the fellaria glacier
col des diablons
the festi glacier
the valle perse
the baltschieder glacier
the grands goulets
mont frety
cima dei gelas
val vajoletti
the glacier de getroz
the col de girardin
passo del mandron
faulhorn
the glacier de bassac
the col de collon
the col des sarrazins
the col de pila
the oetzthaler ferner
the morteratsch glacier
the hufi glacier
the becca di noaschettia
the val prato
becca del merlo
col du lion
the yerva pass
col de nivolet
the pic occidental
the bach alp
the felik glacier
the tuiber pass
col de valloire
the drei schuster spitze
cima del pizzo
valtornanche
the bisi thal
the herbetet
the geschenen thal
gemshorn
vallouise
the fenetre du valsorey
the munster alp
the col de vars
the basse de gerbier
the col de sonadon
the col de la za
the darrei glacier
the bossons glacier
the col del carro
piz bellavista
the como bianco
aiguille noir de peteret
the trogen alp
the innerequell spitz
col de valasco
the col des berches
the defereggen thai
the yal savaranche
the piz ner
the hanig alp
the vallon de la mariande
val di rabbi
aiguille de polset
ville vallouise
the val cairos
the col de yallante
col de la fenetre
the sasaello pass
funffingerspitze
the col de goleon
the pointe de rosablanche
the zwischbergen pass
the becca di luseney
the piz sella
pointe de mary
col de castelnau
pizzo venezia
the col du bonhomme
vallante
the col des courtes
val grosina
the col de rochefort
the fraele pass
the fomo glacier
the tete du toura
the col des masses
the sella pass
piz medel
the glacier de pabeille
the col des grandes murailles
the val de rhymes
the kreuzli pass
the jocelme glacier
the glacier des eivettes
the vallon claus
the col della piccola
val rendena
piz bernina
the yerpeil spitze
the glacier de freboutzie
the gredetsch glacier
the glacier de parste
val malenco
the pisgana pass
the col de valsorey
val di scalve
the piz buin
the col de jaman
the val di gesso
mont colon
the crusehetta pass
the croix de belledonne
the glacier du brouillard
the thaileit spitz
the argentiere glacier
val nambino
the cimes blanches
val federia
val lavinuoz

the col de sageroux
the pic olan
the col de bellevue
col de la croix
the maderaner
tete du salude
rhein valley
the pierre de beranger
val de lys
the val di valasco
piz quatervals
the pic de montandayne
val bendena
corno di dosde
the chardon glacier
the antabbia glacier
the monte di zocca
the becca de guin
the col de mary
the vallon de sellavieille
alpes pennines
the val de hibou
lebendun glacier
the scerscen glacier
the col de creton
the col del merlo
the col de vallante
the glacier du grand
the schmidt kamin
the herbetet glacier
the aiguille de la grande sassifire
the st. elias
the gletscher alp
col du thabor
monte rotta
the becco di menzodi
the staffel alp
the pic de la grave
the maderaner thal
the mont giouberny
the medel glacier
the foppa alp
the piz tschierva
the silvretta glacier
the col du fond
the lauzon glacier
the breche glacier
the peuteret ridge
col de la muande
the col di lago
glacier de casse
the col de la sauce
the pointe de colloney
monte bignone
val antigorio
the clos de la cavalle
val seriana
the vuibez glacier
piz roseo
montanvers
val noana
the sella joch
piz buin
the weingarten glacier
the glacier de bonne pierre
the col de champorcher
becca di nona
the col de sassa
the buffalora pass
the ausser locker spitze
the wengera alp
val maisama
the fassa thai
piz tschierva
the trou de toro
the joch pass
the col de lauteraar
the ausser stellihorn
the noaschetta glacier
the col bardoney
dent parassee
the col de chavancour
the col di lana
val mergoscia
the piz terri
the mur de la cote
val joufirey
the alp zarmine
the col de la pointe de bricolla
the chamois pass
the passo rovano
the happen glacier
the roseg thal
the val champagny
the aiguille de peteret
the val di vallante
col de rioufroid
the col de la gailetta
the gorner grat
val giuf
piz tavrii
the col de vauon pierre
the verpeil spitze
the pic lory
the fresnay glacier
the alpe di goj
the la neuvaz glacier
the plan de la cavalle
mont maudit
the cima bianca
the col bonney
the val de la leisse
the parung pass
the col de sais
the basso di muro
glacier de tetre
the mont pelat
angelus spitze
the pujo glacier
the col des sellettes
wellhorn
pic du midi
the grand chain
the grand colouret
the piz morteratsch
the col de frette
the filar glacier
the pointe des plines
monte gazza
the punta bianca
monte zovo
monte como
the col de martignare
montserrat
the petit coluret
allerhochste spitze
the ginevrie alp
the glacier de la frasse
the pointe des salles
the col ferrex
the kleine windgelle
col de la leisse
the grohmann spitze
the col du palet
col becker
the col de grandcrou
the glacier des ignez
the col de belleface
the rochers rouges
scerscen glacier
the col de berard
la tour ronde
the monch joch
the col de chalance
the val viola poschiavina

the becchi della tribulazione
val imagna
banhorn
piz kesch
the pointe du colloney
the sonadon glacier
pointe de marguerite
the col de la brenva
the val ciamoseretto
the flavona alp
allee blanche
maderanerthal
the val de galambre
the col de pevsque
the becco di mezzodi
the glacier de tabuchet
the langthaler ferner
the glacier du casset
the glacier de gebroulaz
the maloja pass
monte spinale
the glacier de miage
the val mora
bardoney glacier
the gran neiron glacier
monte blanc
the glacier de leschaux
the turtmann ridge
colle campaccio
glacier de thomme
col emile pic
the allerhochste spitze
romanche valley
the glacier de saleinaz
the glacier des ignes
mont roselette
the bosco nero alp
the dent de corjon
the teleccio glacier
the glacier de brenva
glacier de la grande
the aiguille du soreiller
the yaleille glacier
aiguilles de valsorey
the pic des opillous
the col de fenstre
glacier de la casse deserte
the olden alp
the val masino
the grand coluret
becca de ortton
mont brute
the eggen spitze
val marson
the col de verbier
the alp di balme
the alpe du tour
the val vermolera
the piz fliana
val antabbia
the bernetsmatt alp
the glacier du grandcrou sud
the col de la casse deserte
the pic du midi de bigorre
col de rochefort
the tour du grand st. pierre
monte vito
etzthaler ferner
the grand pic
the zocco pass
the piz languard
the tour st. andre
val de cogne
the bertol glacier
the pic jocelme
val narcanello
the monte pian
mont clapier
the corno bianco
piz laschadurelja
the monte prese
col de galambre
mont herbetet
val jouffrey
the cima de jazi
pizzo stella
the uschinen grat
val bajon
the broglio glacier
the nantulon glacier
the cresta agiuza
the col de garin
the piz urlaun
val fassa
val lavaredo
bruneggjoch
fillar pass
val ambies
the vallon de la sausse
the vogel joch
aiguilles de la sausse
the punta giordani
the glacier de marinet
the punta trubinesca
the col de la temple
the seisser alpe
the schwarz glacier
the borzago glacier
col du grand sauvage
the schwartz glacier
the hangend glacier
the kleine zinne
spitzbergen
monte civita
the gran sasso
the col des fours
the val de tinges
the col toumanche
birchfluh pass
col de ceuru
the fiinffinger spitze
the aiguille blanche de peuteret
the col du tour
piz roz
the telleccio glacier
the brenva glacier
monte foscagno
pic des agneaux
piz zupo
the glacier du capucin
the punta foura
the val campaccio
val di fontane
the aiguille du glacier
the val vecchia
ochsenthal
bies joch
balmenhorn
the col du sellar
the val di forzo
the wengern scheideck
the col de cheillon
the hochbalm glacier
the rocher de naye
the col du talon
the glacier d'invergnuon
cimes blanches
the glacier de zardesan
col de tondu
the val cavrein
the col de tignes
the montandevne glacier
ost spitze
the col de corneilla
the becca di nona

the val del zebra
the glacier du giant
the pointe de chamossaire
glacier de la bonne pierre
the pointe haute de mary
the col de la coste rouge
val saviore
the pic tyndall
val travignolo
flavona alp
the pizzo tremoggia
the columbus glacier
the bemetsmatt alp
mont ruan
the gauli glacier
the gornerenthal
the glacier du tacul
the brunni glacier
the val de ribou
the charpoua glacier
the alp kimbianco
the sehalliberg glacier
the col du tour noir
the pointe de la sana
col de valloires
the glacier de rochefort
the glacier du grand mean
the val montjoie
the glacier de charpoua
the colle budden
the glacier de galese
the corno piccolo
val travernanzes
the monte maurigno
cima wilma
the bresciana alp
piz linard
the col de giers
the upper glacier
the aiguille de rousselette
the pic d'otemma
the roccia viva glacier
the tuoi glacier
the cima viola
col de la lavey
the cristall joch
the gliems glacier
furgen glacier
the glacier du tabuchet
the col de mont brute
col de chalance
the rhone valley
the pointe percee
buffalora pass
veglia alp
the aiguille du gouter
the ciardonei glacier
the val del sasso
the alpe granus
the bee du grenier
the val varaita
the bernina scharte
the hohberg glacier
the val di rabbi
the col de pargentine
mont mounier
the gepatsch glacier
the val bavona
the val di roda
the col de plantrin
the zardezan glacier
col de puissaille
the val grisanche
the val maisas
the col de la scigne
val formin
val buona
the glacier de cheillon
the glacier de za de zau
the almageler alp
val teresenga
the becca de nona
monte piano
the blumlis alp
the pic du glacier
the oberhom alp
the porchabella glacier
the predil pass
the schmadri glacier
the mesurina alp
the col de la portetta
the mestia pass
val orsine
the corbassiere glacier
piz languard
the val de cogne
the col de la casse
the aiguille blanche de peteret
the aiguille balmat
brunegghorn
sella pass
the otemma glacier
pic verdonne
the saleinaz glacier
val coumera
cima del largo
the mont fleuri
col des rouies
the val d'entremont
the val briina
the pizzo valgrande
the col du lautaret
the col de baline
cima di piazzi
the pizzo del diavel
the glacier de blaitisre
bee de la sciossa
the val tournanche
combe bremond
the glacier de moiry
the col de monciair
the col de la lavey
the combe froide
cima del ges
the sommet nord
the val sinestra
the grande plaque
the alpe di giovo
the glacier du grand tetret
the tic du glacier
the schalliberg glacier
pointe de sainte anne
the val calanca
the col de monei
the oberalp see
the nuefelgiu pass
the como piccolo
the cavergno glacier
the grand glacier
monte gallina
the maya de bricolla
punta trubinesca
faderjoch
the mont falcon
the pic gaspard
the val des dix
the glacier de broglia
the corno del camoscio
the val lavinuoz
the glacier du geant
the pointe de la yalettaj
pic central
the aiguille de miage
monte di pietit
the col de yalpelline
the cima del lago agnel

monte lifero
the pers glacier
monte cistella alta
the duga pass
monte nero
the liappey alp
the pisole alp
the fum glacier
val codera
val camonica
the becca di moncorve
the rochefort glacier
the tiefenmatten glacier
the suphelle glacier
the baltschieder joch
the val angrogna
tour noir
the monte oliveto
the alp crozlina
doldenhorn
valsenestre
the pic bonvoism
the combe de pierre fendue
the blumlis alp glacier
the col de jallorgues
the col de charnier
the hochbalen glacier
val di mello
the col du vol
monte campo
the val di mello
the hussein thal
the cappella di monte
val malvaglia
the col de mesoncles
the pointe de montandayne
val calanca
val d'algone
the val toumanche
piz pisoc
the petit plateau
the alp gnof
the glacier de la tribulation
the col du celar
the criner furke pass
the col vert
the val orsine
the val de lys
col de lauteraar
the cresta gastaldi
monte kantalaizena
val sparlotsch
ausser stellihorn
the palil glacier
the nantillons glacier
the col de champex
the croix blanche
the maloya pass
the hartley spitze
piz yadret
col vieux
the glacier carra
the urbach thal
the piz medel
the col de terbetet
the lampertsch alp
the russein thal
the pointe de sengies
monte cristallina
the turtman glacier
the pic verdonne
the piz bernina
col de pierre fendue
the passo del diavel
the aiguille du gofitfi
the gabelhom glacier
the col de la lanze
the val maira
the glacier de la charpoua
the konig spitze
the crete de claphouse
the val champey
mont vinaigre
the aiguille des charmoz
the glacier de prou
the schweiben alp
val asinozza
aletsch pass
col de blancien
the zardesan glacier
lower glacier
the punta del broglio
the stufenstein alp
the plachten alp
val masino
monte rosso
the col des ecandies
the ziilerthaler ferner
the col de lauzon
the cevedale pass
the fox glacier
monte faroma
the presena glacier
the yentina glacier
piz borel
the fall glacier
the grand ferrand
the bezingi glacier
the scaletta glacier
becca di noaschetta
the tre coci pass
the pizzo san colombano
the st. gotthard pass
val mora
the viubez glacier
monte cavallo
the augstbord pass
the bietsch joch
the monte perdido
the hohberg pass
piz palii
the col vicentino
the val lavizzara
the glacier de dauva blantz
the aroila glacier
val canzoi
picos de mampodre
the col de la grande luis
furggthal
monte najamola
becca de la liaz
the abberg glacier
the col de la gippiera
the col de la plate des agneaux
the col du says
the pass of hannibal
urbach thal
the hohlicht glacier
mount st. elias
the col de traversette
the tour de st. pierre
the hochste spitze
the mont de saxe
the col du lion
the col di finestre
arbola glacier
the pre de bar glacier
the col de mont rouge
the col de la nouva
the montandeyne glacier
the misauna alp
the pointe de bricolla
the piz eoseg
monte cornacchia
the col de eochefort
the cima de jazzi

val belviso
piz st. michael
the col de tracuit
the glacier carre
the felik joch
the gomerer thal
val di zoldo
the val lavaz
the glacier de lechaud
cola di rienzi
monte sissone
the col de galese
the col de rhymes
the glacier du fond
gepaatsch glacier
the col dolent
val tournanche
the ried glacier
leghorn
felik joch
the grenz glacier
the val petrosciana
col de cretan
the col lombard
col du vallon
piz minger
monte folletto
val bavona
the monte nero
the lenta glacier
val brenta
val chigniulascio
val cabione
the vallon des bancs
val cluozza
scerscen pass
the glacier de cijordnove
mont cheillon
bieshorn
the nagler spitz
fillar joch
the aiguille de la yola
the engstlen alp
linththal
the val jouffrey
the tete blanche
the yuibez glacier
monte leone
the almagel alp
the grand tour
monte pian
vallette
the vin du glacier
the val piantonello
the cristallo pass
the col de lautaret
the pointe du mulinet
the zumstein spitze
val pravitale
the cornera alp
rinderhorn
the grande fourche
monte gristallo
the combe de valeiglia
the huddleston glacier
the val maggia
the pizzo valgrande di valle
the montandayne glacier
col du sirac
the macugnaga thal
the mont perdu
val leventina
the glacier de l'arpont
the dungel glacier
the ost spitze
arpette glen
the val di boda
col des aravis
the rosenlaui glacier
the safien valley
the cedeh glacier
the vieux chaillol
biferten glacier
the moming pass
the est spitze
the col de la muande
the cima cadino
the col des hirondelles
the val de st. marcel
the drei zinnen
the val malenco
the tabaretta thal
val anziei
the buss alp
langkofel glacier
corni di verva
monte durano
the alp kobiei
the ban glacier
the upper brenva glacier
monte matto
montenvers
the zagen glacier
the fedaja pass
egginerhorn
the yallon des etancons
lys-joch
the col di galisia
the val vraita
val di susa
the monte bignone
col des grandes bousses
the aiguille du plan
the brunni alp
the pic de retour
the clos des cavales
the glacier de tronchey
the furgge glacier
monte augusta
the val stura
lauteraarhorn
the atlas mountains
the geiss alp
monte oristallo
the vallon de belleville
the aiguille du midi
the grand pinier
piz murtarol
the val tellina
the val di braulio
the langthaler joch
piz palu
the gredetsch thai
the hasli thal
the col du valsorey
the aiguille de yarens
the col du thabor
becca conge
the nantillon glacier
the glacier des amethystes
the ampezzo valley
val di mel
rizlihorn
the val yaraita
the col di telleccio
the cogne mountains
the crozlina alp
yal grisanche
the chemerno pass
the rocher blanc
the monch jock
col verdonne
the val carapiglia
the arbola glacier
the sulzenan glacier
the grindel alp

val campedelle
the konigs spitze
the aiguille de blaitiere
monte caggio
the glacier de sengie
monte eosa
the flambeau des ecrins
monte ventina
the col du charforon
aiguille du plan
the tschierva glacier
the verra glacier
the petit mont colon
the ampezzo dolomites
the val fiera
the corno di dosde
the dossradond pass
martinsloch
val di prato
the monei glacier
the loffel spitz
mont emilius
piz formin
val di lucerna
the col de joux
the col di tenda
val ombretta
glacier des ecrins
the val devero
the ueschinen thal
the noasehetta glacier
the bies glacier
the brunni grat
glacier du clot des cavales
the gasteren thal
the pont de mauvoisin
col paganini
val fonda
the breithom glacier
val vermiglio
the col du midi
valtellina
the pointe de garin
the matscher thai
mont pelat
the pizzo campo tenca
sckontauf spitze
augstbord pass
the val lazin
the glacier de zardezan
the bee de phomme
zocca pass
the glacier du bouchet
the val de viu
the col de bardoney
monte forato
val challant
val ruvinian
the val pradidali
the pointe de tinneverges
the zmutt ridge
the bies joch

Table des matières

MOTS-CLÉS : oronyme, NER, apprentissage automatique, TAL, BERT

RÉSUMÉ

Dans cette tâche, nous avons élaboré un jeu de données en utilisant 1221 oronymes pour affiner les modèles avancés basés sur le Transformer tels que BERT. L'objectif principal est d'identifier les oronymes à partir de notre corpus spécialisé contenant 1,4 million de mots de récits d'explorations des Alpes. Nous avons ensuite utilisé ce jeu de données pour évaluer la performance de spaCy, un modèle général. Notre ambition est d'améliorer la précision de la reconnaissance pour les futurs projets de dénomination géophysique. De plus, nous souhaitons contribuer à l'étude et à l'application du traitement automatique de la langue, en particulier dans le domaine de la reconnaissance des entités nommées.

KEYWORDS : oronym, NER, machine learning, NLP, BERT

ABSTRACT

In this task, we created a dataset using 1221 oronyms to fine-tune advanced Transformer-based models such as BERT, with the aim of identifying oronyms from our specialised corpus of 1.4 million words of Alps exploration narratives. The resultant dataset was also employed to evaluate spaCy, a general-purpose model. We strive to enhance the precision of identification for forthcoming projects in geophysical naming, as well as to support the study and application of natural language processing, particularly in the field of named entity recognition.