

## Avant-propos

La lecture du présent rapport ne nécessite pas de connaissance préalable en traitement automatique de la langue, en apprentissage automatique ou en apprentissage profond.

Il est cependant préférable d'avoir quelques connaissances en informatique pour comprendre les enjeux du stage.

Par ailleurs, de nombreux termes techniques sont utilisés. Bien qu'habituellement rencontrés sous leur forme anglaise dans le domaine, ces termes ont été traduits en français, à l'exception de certains sigles utilisés tels qu'ils dans la littérature française.

Tous les termes techniques seront donc introduits en français, accompagnés d'une explication, de la traduction anglaise et du sigle anglais si nécessaire. De plus, une partie du rapport est dédiée à l'explication des termes et concepts utilisés, et un glossaire est présent en fin d'ouvrage. Enfin, quelques notes de bas de page fournissent des précisions ponctuelles sur certains concepts utilisés. Elles sont notées en exposant.

Les références aux sources sont marquées entre crochets, et numérotées par ordre d'utilisation.

Ce rapport a été réalisé avec  $\text{\LaTeX}$  et les diagrammes présentés ont été produits par nos soins avec l'outil TikZ, ce qui explique l'absence de source pour les figures.

# Table des annexes

<b>A</b>	<b>Rapports d'avancement du projet GMSNN</b>	<b>87</b>
A.1	Informations sur les rapports contenus dans la présente annexe	87
A.2	Étude des problèmes de mémoire	88
A.3	Sauvegarde du modèle	93
A.4	Solution tentative pour les fuites mémoires	97
A.5	Problèmes théoriques liés à l'entraînement batch par batch	99
A.6	Tentative de réduction du temps de calcul par utilisation de l'algorithme d'entraînement « <i>Truncated BPTT</i> »	102
A.7	Entraînement couche par couche	104
A.8	Premier test du modèle réimplémenté	106
A.9	Premier entraînement complet du modèle réimplémenté	107
A.10	Premier entraînement à 50 époques du modèle réimplémenté	109
A.11	Premier test du modèle multi-échelles	113
A.12	Comparaison des stratégies de fusion des résultats des différentes couches	117
A.13	Test des effets du changement de taille des paquets ( <i>batch</i> )	122
A.14	Entraînement sur le corpus complet avec beaucoup de temps alloué	124
A.15	Changement de stratégie de gestion d'historique	131
A.16	Test de l'implémentation des <i>batches</i>	138
A.17	Test des performances des <i>batches</i>	142
A.18	Test des performances des <i>batches</i> sur 50 époques	146
A.19	Test des performances des différentes améliorations	151

<b>B</b>	<b>Rapports d'avancement du projet PAPUD</b>	<b>159</b>
B.1	Informations sur les rapports contenus dans la présente annexe	159
B.2	Informations générales	160
B.3	Résultats de l'implémentation basique	162
B.4	Paquets ( <i>batches</i> ) simultanés	164
B.5	Analyse du pic de performance	168
B.6	Rapport de la réunion avec les autres membres du projet	172
B.7	Optimisation du taux d'apprentissage	174
B.8	Effets de l'optimisation du taux d'apprentissage	178
B.9	Taille de <i>batch</i> binaire	180
B.10	Performances du lecteur de corpus multi-fichiers multi-processus	182
<b>C</b>	<b>Copie de la convention de stage</b>	<b>187</b>
<b>D</b>	<b>Copie de l'avenant à la convention de stage</b>	<b>193</b>

# **A Rapports d'avancement du projet GMSNN**

## **A.1 Informations sur les rapports contenus dans la présente annexe**

Les sections suivantes contiennent les rapports intermédiaires fournis au maître de stage tout au long du projet GMSNN.

### **A.1.1 Format d'origine des rapports**

Le langage Markdown, plus spécifiquement dans le dialecte nommé Gitlab Flavoured Markdown (littéralement « Markdown à la saveur de Gitlab »), fournit une syntaxe facile à lire et à écrire. Il permet la rédaction de documents agrémentés entre autres d'images, de formules, de tableaux et de fragments de codes. Enfin, l'affichage du Gitlab Flavoured Markdown est supporté par Gitlab.

Ces particularités en font un langage de premier choix pour l'écriture de rapports destinés à être lus au format informatique directement sur Gitlab.

### **A.1.2 Transcription des rapports**

L'intégration des rapports intermédiaires dans ce rapport à nécessité l'adaptation au format papier du contenu rédigé en Gitlab Flavoured Markdown.

Certains éléments n'ont pas pu être transcrit de façon exacte, en particulier les liens et les tableaux et images de grand taille. Ces éléments ont donc été adaptés.

### **A.1.3 Contenu et langue des rapports**

Le contenu des rapports n'a été ni modifié ni corrigé, et est livré en anglais tel qu'écrit à l'origine.

L'anglais a été choisi comme langue de rédaction des rapports pour maintenir la cohérence avec le code, écrit et documenté en anglais lui aussi, et avec la littérature, principalement rédigée en anglais. Ce choix évite aussi d'alourdir le contenu déjà complexe des documents avec des traductions maladroites.

## **B Rapports d'avancement du projet PAPUD**

### **B.1 Informations sur les rapports contenus dans la présente annexe**

Les sections suivantes contiennent les rapports intermédiaires fournis aux membres de l'équipe de projet au cours du projet PAPUD.

#### **B.1.1 Format d'origine, transcription et contenu des rapports**

Pour les mêmes raisons que pour l'annexe précédente (décrite dans la section A.1, page 87), les documents présentés dans cette annexe ont été rédigés en anglais, au format Gitlab Flavoured Markdown; les versions présentées ici sont des transcriptions aussi fidèles que possible de ces documents.