Facultés Universitaires Notre-Dame de la Paix

INFO M210 - STAGE

- Rapport de Stage -

Laurent JAKUBINA
ljakubin@student.fundp.ac.be

Promoteur de Mémoire:
Naji HABRA
FUNDP

Maitre de Stage:
Houari SAHRAOUI
Philippe LANGLAIS
Université de Montréal

1 Mars 2012

Table des matières

1 Introduction

Ce rapport de stage a pour but de résumer et d'expliquer brièvement le déroulement de mon stage effectué dans le cadre de ma deuxième maitrise, dernière année de mon cursus d'études des sciences informatiques aux Facultés Universitaires Notre Dame de la Paix à Namur. Étroitement lié à mon mémoire supervisé par mon promoteur et professeur Naji Habra, mon stage a eu lieu à l'Université de Montréal, plus précisément au Département d'Informatique et Recherche Opérationnelle ou DIRO. J'y ai travaillé sous la tutelle du professeur Houari Sahraoui, membre de l'équipe GEODES (génie logiciel), et de Philippe Langlais, membre de l'équipe RALI (ingénierie linguistique). Mon travail a consisté à aborder l'idée de « Génération automatique de commentaires en utilisant les techniques de la traduction automatique statistique ». Je commencerai donc par expliciter un peu plus le sujet ainsi que les objectifs. Ensuite, je déploierai un paragraphe sur le déroulement du stage, suivi d'explications plus poussées sur certaines activités et développements effectués durant le stage. Et finalement je terminerai par une conclusion synthétisant mon rapport au stage.

2 Objectifs

2.1 Énoncé

Voici l'énoncé tel qu'il m'a été proposé :

« Le sujet consiste à étudier la possibilité de générer des commentaires pour des éléments de programmes, tels que les méthodes à partir de corpus de données (paires code-commentaire). Pour ce faire, nous pensons à l'utilisation des techniques de traduction statistiques qui utilisent une base d'exemples (paire de segments de texte équivalents de deux langues). Ces techniques de traduction donnent des résultats impressionnants (voir par exemple les nouveaux traducteurs de Google). Le projet consistera dans une première phase à créer un corpus de données (méthodes et leurs commentaires) à partir des logiciels open source. La deuxième phase consiste à adapter des traducteurs existants à la problématique de la génération des commentaires. »

2.2 Définition des objectifs

Après une première réunion d'introduction et d'explicitation du contexte, l'objectif principal était clair : étudier la possibilité de générer automatiquement des commentaires à partir des techniques de traduction automatique statistique, et plus précisément des techniques d'entrainement de corpus de textes traduits mis en parallèles.

Pour ce faire, une série d'étapes-avancements se sont dessinés/s'est dessinée (? ? ?) :

- S'approprier le contexte des techniques de traduction statistique : corpus de textes traduits mis en parallèle à entrainer afin de générer des traducteurs.
- Se familiariser avec l'outil « Moses », logiciel open source d'entrainement de corpus, notamment en lisant son manuel d'utilisateur.
- S'intéresser aux outils proposés par une étudiante du laboratoire GEODES concernant ses travaux sur les commentaires.
- Préparer un corpus d'entrainement et envisager de le préprocesser.
- Entrainer le corpus, évaluer les résultats et préparer une validation.

Dans la suite du rapport, chaque étape sera d'abord située dans le temps grâce au semainier et expliqué plus en détails dans la section sur les activités scientifiques et les développements.

3 Déroulement du stage

3.1 Semainier

Afin d'avoir en premier lieu une vue globale du déroulement du stage, j'ai décidé de commencer par le semainier. Pour chaque semaine j'y explique brièvement ce que j'ai fait. Il faut se rapporter à la section 4 de mon rapport de stage pour avoir plus de détails sur certaines des activités effectuées. Notons que toutes les lectures effectuées durant le stage seront explicitement nommées dans la bibliographie située à la fin de ce rapport.

- *Semaine 1 : 5 Septembre au 9 Septembre*

  Découverte de la littérature existante autour des thèmes de « Traduction Automatique Statistique » et « Génération Automatique de commentaires » : Lectures d'articles et mise en contexte.

- *Semaine 2 : 12 Septembre au 16 Septembre*

  Découverte de Moses [1], logiciel de traduction automatique statistique permettant l'entrainement de corpus de textes : installation et lecture de son manuel d'utilisateur.

- *Semaine 3 : 19 Septembre au 23 Septembre*

  Suite et fin de la lecture du manuel d'utilisateur de Moses. Lecture rapide de différents mémoires sur la traduction automatique statistique.

- *Semaine 4 : 26 Septembre au 30 Septembre*

  Lecture du mémoire d'une étudiante en maitrise du laboratoire GEODES sur l'extraction des commentaires de classes Java. Découverte de ses outils et des outils suivants : Le parseur SableCC et la librairie JDom.

- *Semaine 5 : 3 Octobre au 7 Octobre*

  Compréhension et intégration des outils développés autour de l'extraction des commentaires dans un projet Java. Modification des outils et implémentation visant à préparer le corpus de commentaires. Apprentissage du langage de requêtes XPATH et des expressions régulières en Java (regex).

- *Semaine 6 : 10 Octobre au 14 Octobre*

  Implémentation de l'extraction des commentaires et du code lié à ceux-ci, enregistré dans deux fichiers textes séparés à fin de préparer un bitexte commentaires/code parallélisé. Identification des différents types de commentaires.

- *Semaine 7 : 17 Octobre au 21 Octobre*

  Implémentation de la récupération du code du corps d'une méthode ou juste de son entête. Choix d'utiliser le logiciel open source JHotDraw, défini comme un cas d'étude, bien documenté, ... pour effectuer les premiers tests.

- *Semaine 8 : 24 Octobre au 28 Octobre*

  Premier choix de préprocessing du corpus : implémentations de nouvelles fonctionnalités. Correction de bugs et optimisations des fonctionnalités.
  Préparation du corpus, entrainement du corpus , premières batteries de tests de traduc-

1. http ://www.statmt.org/moses/

tion, évaluation et commentaires.

- *Semaine 9 : 31 Novembre au 4 Novembre*

  Choix de préprocessing du corpus : implémentations de nouvelles fonctionnalités. Adaptation du corpus, entrainement du corpus modifié, nouvelles batteries de tests de traduction, évaluation et commentaires.
  Implémentation du splitter d'identifiants.

- *Semaine 10 : 7 Novembre au 11 Novembre*

  Adaptation du corpus (splitter d'identifiants), entrainement du corpus modifié, nouvelles batteries de tests de traduction, évaluation et commentaires.

- *Semaine 11 : 14 Novembre au 18 Novembre*

  Correction de bugs et optimisations des fonctionnalités.
  Adaptation du corpus, entrainement du corpus modifié, nouvelles batteries de tests de traduction, évaluation et commentaires.
  Choix de compléter le corpus avec toutes les versions de JHotDraw.

- *Semaine 12 : 21 Novembre au 25 Novembre*

  Téléchargement de toutes les versions de JHotDraw et extraction de leurs codes et commentaires à l'aide de mes outils. Entrainement du corpus agrandi, nouvelles batteries de tests de traduction, évaluation et commentaires.
  Implémentation d'un script d'automatisation de traductions.
  Choix d'essayer de générer des commentaires pour une version récente de JHotDraw à partir des anciennes versions.

- *Semaine 13 : 28 Novembre au 2 Décembre*

  Adaptation du corpus en fonction du dernier choix, entrainement du corpus modifié, nouvelles batteries de tests de traduction, évaluation et commentaires.
  Choix d'agrandir au maximum le corpus à l'aide d'autres projets.

- *Semaine 14 : 5 Décembre au 9 Décembre*

  Téléchargement d'une vingtaine de projets Java open source et extraction de leurs codes et commentaires à l'aide de mes outils. Entrainement du corpus agrandi, nouvelles batteries de tests de traduction, évaluation et commentaires.

- *Semaine 15 : 12 Décembre au 16 Décembre*

  Présentation de fin de stage et discussion de l'avancement effectué sur les 15 semaines.

3.2 Encadrement

L'environnement et les conditions de travail ayant aussi leurs influences sur la productivité, je décris ici le milieu dans lequel j'ai travaillé pendant mes trois mois et demi de stage. Je commencerai par présenter un peu le campus de l'université de Montréal, ensuite le laboratoire de Génie Logiciel (GEODES) et finalement, la relations avec mes maitres de stage.

3.2.1 L'université de Montréal

Le campus de l'Université de Montréal est immense. Situé sur le Mont-Royal, l'Udem est reconnaissable de loin grâce à son pavillon Roger Gaudry et est un des symboles de la métropole. Sur le campus, toutes les nationalités s'y promènent, tel qu'il y est rare d'y croiser deux fois la même personne. Ce qui est sûr, c'est que l'université de Montréal mérite bien sa renommée internationale et cela se ressent quand on se promène sur le campus.

3.2.2 Le laboratoire GEODES et l'équipe

Situé dans le département de d'informatique et de recherche opérationnel (DIRO) de la faculté des Arts et des Sciences, le laboratoire de Génie Logiciel permet à une équipe de 10 à 15 personnes de travailler soit dans le cadre de sa maitrise, soit dans le cadre de son doctorat. Des machines Windows et Linux sont mises à dispositions et aussi, depuis récemment, un serveur de calcul qui, notamment, m'aura aidé durant mes entrainements sur des corpus de grande taille.

L'équipe travaillant au laboratoire, supervisée par Houari, est très sympa et multiculturelle. Nous avons notamment organisé des repas entre nous pour apprendre à mieux nous connaître, ce qui a donné lieu à une bonne ambiance durant les dures journées de labeurs.

Et pour terminer, chaque jeudi matin était réservé afin que l'équipe se mette autour de la table et discute de sujets concernant le laboratoire. On assistait aussi à au moins une présentation d'un étudiant du laboratoire, la plupart du temps dans le cadre d'une préparation d'une présentation pour l'étudiant en question.

3.2.3 Les promoteurs

Mes promoteurs respectifs, Houari et Philippe, se sont montrés très disponibles, que cela soit par mail ou en allant les chercher directement dans leur bureau. Cela a été un plaisir de travailler avec eux.
Notons aussi que Houari, superviseur du laboratoire GEODES, passait presque tous les jours afin d'avoir les dernières nouvelles sur l'avancement du travail.
Aussi, on essayait avec Houari et Philippe, de faire au moins une réunion par semaine afin d'évaluer les progrès effectués ainsi que déterminer les prochaines étapes du développement.

3.2.4 Les « à cotés »

Par l'expression les « à cotés », je parle évidemment des possibilités d'expériences en dehors du travail qu'il était possible de vivre en faisant son stage à Montréal. Et en effet, il n'y avait pas moyens d'être déçu. Montréal, plus qu'une ville, une métropole, possédant plusieurs visages, telle que chacun peut y trouver un endroit où il se sent chez lui. Ces trois mois et demi passés là-bas m'ont permis de bien cerner la mentalité montréalaise mais aussi de goûter un peu à la culture québécoise, notamment en faisant quelques voyages en dehors de la métropole. Je ne détaillerai pas chaque voyage mais je terminerai en ajoutant que la possibilité de faire un stage à l'étranger ne fait que renforcer l'expérience du stage.

3.3 Appréciation

Pour terminer cette section sur le déroulement du stage et avant d'expliquer plus en détails les activités scientifiques et les développements effectués durant le stage, voici en quelques phrases, mon appréciation résumée vis à vis de l'expérience unique qu'à été de faire mon stage de deuxième maitrise à Montréal, et plus précisément au laboratoire GEODES de l'Université

de Montréal. En plus d'être un magnifique voyage, on se sent sortir de ce moment avec un gain de maturité mais aussi d'expériences de travail. L'approche « Ouverture d'esprit » promue par l'éducation universitaire en est à son apogée avec un stage de cette ampleur. Et finalement, c'est parfois l'occasion de faire de magnifiques rencontres...

4 Activités Scientifiques et développements

Durant le semainier, j'ai rapidement parlé de toutes les tâches que j'ai effectuée durant le stage. En effet, celle-ci n'ont pas été détaillée, ce que je compte faire avec cette section 4.

4.1 Appropriation du contexte

Avant de rentrer dans le vif du sujet, autrement dit de générer automatiquement des commentaires à l'aide de la traduction automatique statistique, il m'a fallu passer un peu en revue la littérature existante sur ces deux sujets. Pour cela j'ai lu un certains nombres de papiers et de mémoires provenant soit de mes promoteurs soit d'Internet. Touchant à l'ingénierie linguistique pour la première fois, j'ai appris aussi les notions de base à l'aide de cours trouvés sur internet. Et finalement, afin de me préparer à l'entrainement de corpus de bitexte, il m'a été donné de lire le manuel du logiciel open source Moses et de procéder à son installation. Dans la foulée, j'ai réalisé quelques entrainements sur des petits corpus donnés dans les tutoriels du manuel. La liste des références situées à la fin de ce rapport reprend l'ensemble des titres des lectures que j'ai effectuée.

4.2 Découverte des outils

Le premier outil auquel j'ai été confronté, déjà cité dans le paragraphe précédent, est Moses. Disponible gratuitement sur http ://www.statmt.org/moses/, ce logiciel open source permet de réaliser des entrainements de corpus de textes bilingues afin de générer des modèles de traduction. Afin de préparer un corpus de textes, qui dans ce cas-ci rappelons-le, est d'un côté, le code et de l'autre, les commentaires, il me fallait un moyen d'extraire ceux-ci d'un projet Java. C'est en lisant le mémoire d'une étudiante du laboratoire GEODES que j'ai trouvé les outils adéquats. Premièrement, SableCC qui est un générateur de compilateur/interpréteur en Java. Celui-ci permet notamment de représenter un programme java sous forme d'un arbre en XML. Ensuite, en utilisant (J)DOM pour manipuler ces fichiers XML ainsi que le langage de requêtes XPATH pour effectuer des actions précises sur les nœuds de l'arbre, il m'était possible de localiser et extraire les commentaires ainsi que le code lié à ceux-ci. Finalement, afin d'optimiser l'ensemble de mes requêtes, j'ai appris à utiliser au mieux les expressions régulières.

4.3 Implémentations

A partir de ces outils, l'implémentation a donc consisté à coder des fonctionnalités dans le langage Java permettant d'extraire les commentaires et le code lié à ceux-ci d'une part, mais aussi d'autre part, d'avoir une certaine aisance à retravailler les fonctionnalités pour permettre de répondre au besoin de préprocesser les corpus en fonction des expériences qu'il serait nécessaire d'effectuer. C'est donc une série de fonctionnalités que j'ai créées, modifiées et adaptées dans le temps en plusieurs versions :

- Détecter de quel type est un commentaire (de classe, de méthode ou autres) et pouvoir choisir quel type de commentaire on traite.

– Récupérer le code lié aux commentaires sélectionnés.
– Analyser et traiter les caractères spéciaux afin de d'avoir un corpus propre.
– Splitter d'identifiants (couper un nom de méthode en fonction des majuscules).
– Etc.

4.4 Préparations du corpus

Avant de se lancer dans l'entrainement, il me fallait évidemment du contenu. Contenu sur lequel j'allais appliquer mes outils afin de préparer un corpus de bitexte code-commentaire. Ce contenu, c'est un ou des programmes Java desquels on va extraire le code et les commentaires. Conseillé par Houari, car connu pour respecter certains « bons » principes de codage et notamment d'être bien documenté (cas d'université), c'est le logiciel open source JHotDraw qui m'a servi de base pour les premières expériences d'entrainements et de tests d'entrainements. Par la suite et comme déjà noté dans le semainier, j'ai dû télécharger et traiter une vingtaine de projet open source différents afin d'agrandir le corpus.

4.5 Entrainements et Tests de Traductions

Une fois le corpus créé à partir de l'extraction des commentaires et du code d'un programme Java, il est temps de lancer l'entrainement à l'aide de Moses sur le bitexte. Cette manipulation va consister à spécifier à Moses en entrée, le bitexte code – commentaire ainsi qu'un certains nombres de paramètres de configuration. Durant l'entrainement, le logiciel calcule les probabilités qu'un certain fragment de texte se répète en parallèle dans sa langue source et dans sa langue cible. C'est ce qu'on appelle la traduction statistique, que j'expliquerai plus dans mon mémoire. L'entrainement est un processus long en fonction de la taille du corpus. Pour donner un exemple, avec des corpus de plusieurs millions de lignes, c'est des heures que prend le logiciel pour effectuer les calculs. En sortie, le logiciel génère un certain nombre de fichiers qui vont être passés en paramètre de la commande de traduction. Cette commande, on l'appelle ensuite sur le texte que l'on souhaite traduire et en output, on reçoit le texte traduit. Le processus est fonctionnel, ce qui nous reste à évaluer, c'est la qualité de la traduction (génération du commentaire).

4.6 Évaluations

Comme déjà dit dans le paragraphe précédent, ce que l'on cherche à évaluer, c'est la qualité de la traduction. Pour cela, j'ai constitué un ensemble de 100 morceaux de code extraits du corpus sur lequel on est entrain de travailler. Je les aie ensuite « traduits » avec le traducteur résultant de l'entrainement effectué avec le dit-corpus. Je me retrouve alors avec un ensemble de 100 commentaires venant de la traduction, que je suis capable d'évaluer en les comparant avec les 100 commentaires originaux liés aux bouts de code choisis. En définissant une métrique telle que par exemple, si le commentaire généré permet de transmettre le sens exact du commentaire original alors on attribue une valeur « 3 ». Si par contre, le commentaire généré ne permet pas de comprendre le sens du bout de code ou dans le pire de cas, ne représente aucune valeur, alors on attribue une valeur « 0 ». A l'aide de cette échelle de valeur, il est à la fois simple et efficace d'évaluer les traductions générées et d'émettre, comme on le cherchait, un avis sur l'idée de générer des commentaires automatique grâce à la traduction automatique statistique.

4.7 Résultats et Conclusions

Après avoir effectué une série d'entrainements de bitextes de différentes caractéristiques, ainsi que des batteries de tests de traduction code - commentaire que j'ai pu évaluer, les résultats

sont sans appel : Oui la technique de la traduction automatique statistique appliquée à un bitexte code - commentaire est fonctionnelle, en effet vu que le principe a été étudié pour calculer des probabilités d'apparitions de série de caractères, sans aucun rapport avec les langues écrites dans le corpus. Mais par contre, les « commentaires » générés, s'il on peut encore appeler cela des commentaires, sont de valeurs nulles. Les « commentaires » générés se retrouvent être une série de mots sans aucun lien entre eux, sans aucun sens,... Cependant, de nombreuses améliorations sont encore applicables : préprocesser le corpus selon d'autres idées (retirer les mots clés, balises particulières, adapter la taille des phrases du corpus, ...), modifier le logiciel d'entraînement (ex : Moses), euristiques, etc.

5 Conclusion

La conclusion finale de ce rapport ne va pas être longue, et pour cause, la section 3.3 Appréciation apporte déjà une conclusion concernant le déroulement global du stage et la section 4.7 Résultats et Conclusions donnent les conclusions des résultats des expériences effectuées. J'inviterai donc le lecteur à aller lire ces deux points afin de « compléter » cette conclusion. Mais pour finir, je dirai que les objectifs ont été atteins. L'idée d'appliquer la traduction automatique statistique sur un corpus code-commentaire afin de générer des commentaires a été fonctionnelle dans un sens. Les résultats sont décevants sur la qualité de la traduction évidemment, mais l'idée est viable et peut encore être étudiée et réfléchie sur de nombreux points. J'espère donc que la recherche va continuer sur cette voie, car en effet, la génération automatique de commentaires, n'est-elle pas pour tout programmeur qui se respecte, un rêve ? ...