Concours National d'Informatique

Sujet de demi-finale Paris 3

7 mars 2009

Traduction

1 Préambule

Bienvenue à **Prologin**. Ce sujet est l'épreuve écrite d'algorithmique et constitue la première des trois parties de votre demi-finale. Sa durée est de 3 heures. Par la suite, vous passerez un entretien (20 minutes) et une épreuve de programmation sur machine (4 heures).

Ceci est une épreuve d'algorithmique. Cela signifie que ce qui intéresse les correcteurs n'est pas la manière dont vous écrivez votre code, qui sera testée cet après-midi, mais votre manière de réfléchir et de résoudre des problèmes.

À ce titre, tous les langages sont autorisés, y compris le pseudo-code, pourvu que vous indiquiez lequel vous utilisez. Si vous éprouvez une quelconque difficulté avec votre langage, vous pouvez ainsi expliquer « en français » votre manière de résoudre la question, à condition que vous indiquiez un processus détaillé facilement transposable en un programme.

Conseils

- Lisez bien tout le sujet avant de commencer.
- **Soignez la présentation** de votre copie.
- N'hésitez pas à poser des questions.
- Si vous avez fini en avance, relisez bien.
- N'oubliez pas de passer une bonne journée.

Remarques

- Le barème est donné à titre indicatif uniquement.
- Indiquez lisiblement vos nom et prénom, la ville où vous passez la demi-finale et la date, en haut de votre copie.
- Si vous trouvez le sujet trop simple, relisez-le, réfléchissez bien, puis dites-le nous, nous pouvons ajouter des questions plus difficiles.
- Le barème récompense les algorithmes les plus efficaces : écrivez des fonctions qui trouvent la solution le plus rapidement possible.
- Ce sont des humains qui lisent vos copies : laissez une marge, aérez votre code, ajoutez des commentaires (**seulement** lorsqu'ils sont nécessaires) et évitez au maximum les fautes d'orthographe.

2 Sujet

Introduction

Nous allons aujourd'hui nous intéresser à la construction d'un système de traduction automatique de texte depuis une langue vers une autre. Pour cela nous mettrons de côté certaines choses, notamment les flexions (transformation d'un mot suivant le contexte grammatical : conjugaisons, accord, déclinaisons...). Nous tiendrons en revanche compte du fait que tous les langages n'expriment pas les concepts de la même façon.

Question 1 (1 point)

Supposons que nous disposions d'un dictionnaire simplifié d'une langue vers une autre : à chaque mot de la langue de départ sont associés un ou plusieurs mots de la langue d'arrivée. Un indice de pertinence permet de savoir s'il s'agit exactement du même concept ou d'un concept approchant.

Proposez une structure de données permettant d'implémenter ce dictionnaire, ainsi qu'une fonction permettant de connaître la meilleure traduction d'un mot donné.

Question 2 (1 point)

Au sein d'une phrase, les fonctions syntaxiques que peut prendre un mot sont nombreuses : pronom, nom, adjectif, verbe, adverbe... On parle de *nature*.

Selon le contexte, un même mot peut d'ailleurs avoir une fonction différente. Par exemple le mot français « orange » peut être un nom commun (le fruit) ou bien un adjectif (la couleur). Lors de la traduction on veut bien entendu un mot de même nature à l'arrivée. Ainsi dans un dictionnaire les natures sont séparées, et souvent distinguées par un nombre par exemple.

D'autre part les langues ont souvent plusieurs registres, depuis le soutenu jusqu'au familier,

les différences pouvant être subtiles [1].

Question 2a

Proposez une structure de données pour décrire la nature et le registre de langue d'un mot.

Question 2b

Modifiez la structure de données du dictionnaire pour permettre de connaître la nature ou le niveau de langage des mots et de distinguer deux mots orthographiés de la même manière mais dont la nature est différente.

Question 3 (4 points)

Une phrase est composée d'ensembles de mots remplissant des fonctions telles que sujet, verbe, complément. Plus généralement, à l'échelle d'une phrase, on parle de *syntagmes* : syntagme pronominal, syntagme nominal, syntagme adjectival, syntagme verbal, syntagme adverbial. Un syntagme peut lui même être englobé dans un syntagme supérieur, et inversement inclure des syntagmes inférieurs.

Par exemple la phrase « Des étudiants venus de partout en France se réuniront au mois d'avril. » se décompose en un syntagme nominal « des étudiants venus de partout en France » et un syntagme verbal « se réuniront au mois d'avril », qui lui-même se décompose en un syntagme verbal « se réuniront » et un syntagme nominal « au mois d'avril », ainsi de suite, jusqu'à arriver aux mots.

Finalement, un syntagme est quelque chose qui contient soit un mot, soit d'autres syntagmes, et est caractérisé par sa fonction.

Question 3a

Proposez une structure de données permettant d'exprimer une phrase suivant une telle décomposition.

Question 3b

On vous donne une phrase décrite avec cette structure. Pour chaque mot de la phrase, indiquez la meilleure traduction dont vous disposez du point de vue de la pertinence, ou à défaut, qu'aucune traduction n'est disponible.

1. Certaines langues possèdent également les notions de respect et d'humilité, et permettent de les exprimer distinctement, mais nous n'en tiendrons pas compte ici.

Question 3c

Tant que vous y êtes, ce pourrait être amusant de tester les autres traductions. Pour une phrase donnée, renvoyez toutes les combinaisons de traductions possibles d'après votre dictionnaire.

Question 4 (4 points)

À ce stade, nous avons fait de la traduction mot à mot. Bien entendu, si c'était aussi simple, ça se saurait. :-) Il est donc probable que les résultats de la question précédente ne valent guère mieux que ceux donnés par un quelconque traducteur automatique gratuit sur Internet (*question subsidiaire* : pourquoi Babelfish s'appelle-t-il Babelfish ?).

La grammaire dépend en effet de la langue, et exprimer une même idée peut donc se faire de façon très différente d'une langue à l'autre. Nous allons maintenant tâcher de construire des phrases correctes.

Pour cela nous allons nous doter d'un dictionnaire qui ne se contente pas de donner les traductions d'un mot vers un autre, mais d'une construction vers une ou plusieurs autres (avec toujours un indice de pertinence et un autre de registre de langue). La traduction d'un mot vers un autre ne sera alors plus qu'un cas particulier de construction composée d'un seul mot et dont la traduction est également une construction composée d'un seul mot.

Dans un premier temps nous allons permettre d'avoir dans notre dictionnaire des expressions toutes faites. Par exemple la phrase « Ça ne peut pas être aidé. » est certes la traduction littérale exacte de l'anglais « It cannot be helped. », mais certainement pas une traduction correcte.

Question 4a

Proposez en conséquence une nouvelle structure de données pour notre dictionnaire, qui permette d'y insérer non plus seulement des mots, mais aussi une expression et sa traduction.

Question 4b

Supposons que vous ayez une phrase en mémoire, représentée grâce à votre structure de données, et que vous ayez dans votre dictionnaire le proverbe « qui vole un œuf vole un bœuf ». C'est un syntagme verbal, qui se décompose en un syntagme pronominal « qui vole un œuf » et un syntagme verbal « vole un bœuf ». Le premier se décompose ensuite en un syntagme pronominal « qui » et un syntagme verbal « vole un œuf ».

Indiquez si cette phrase contient au moins un syntagme verbal contenant lui même un syntagme pronominal et un syntagme verbal.

Question 4c

Indiquez si le syntagme pronominal se décompose en un syntagme pronominal contenant le pronom « qui » et un syntagme verbal.

Question 4d

Écrivez maintenant une fonction permettant de déterminer si une expression arbitraire présente dans le dictionnaire est présente dans la phrase.

Question 5 (4 points)

Notre système de traduction s'améliore, mais ce n'est pas encore ça. Nous savons désormais traduire des mots, des expressions toutes faites, mais au final ça reste du mot à mot. Dans un second temps nous allons donc encore améliorer notre dictionnaire pour pouvoir faire des manipulations plus complexes.

Prenons l'exemple de la phrase « Les candidats galèrent car ils ont oublié leurs lointains cours de grammaire. ». Elle est composée de deux syntagmes verbaux « les candidats galèrent » et « ils ont oublié leurs lointains cours de grammaire », articulés autour de la conjonction « car ». Le premier exprime la conséquence, le second, la cause.

La phrase « Comme ils ont oubliés leurs lointains cours de grammaire, les candidats galèrent. » exprime une idée proche, mais en insistant sur la cause lorsque la précédente insistait sur la conséquence. En français on exprime cette différence en changeant de conjonction, ce qui modifie tout l'ordre de la phrase. Dans d'autres langues, il se pourrait que l'ordre soit encore complètement différent. Il nous faut donc pouvoir exprimer cet ordre dans lequel vont se retrouver cause, conséquence, et conjonction.

En résumé, on veut donc savoir comment exprimer dans une autre langue *(syntagme verbal conséquence) (conjonction) (syntagme verbal cause)*, ce qui implique que dans notre dictionnaire nous allons vouloir mettre des choses telles que *(syntagme verbal conséquence) (conjonction « car ») (syntagme verbal cause)* ou encore *(conjonction « comme ») (syntagme verbal cause) (syntagme verbal conséquence)*, dans les deux langues bien sûr.

Question 5a

Proposez tout d'abord une structure de données pour exprimer une telle construction au sens général, qui n'indique plus nécessairement des mots, mais plutôt des contraintes.

Question 5b

L'ordre des syntagmes étant potentiellement différent à l'arrivée, il faut pouvoir les identifier. Indiquer *cause* ou *conséquence* n'est peut-être pas une bonne idée, car on ne peut pas être

exhaustif sur toutes les constructions pouvant exister (c'est au dictionnaire d'être exhaustif, pas au code).

Proposez donc un moyen d'identifier la correspondance entre les syntagmes dans la langue de départ et d'arrivée.

Question 5c

Modifiez en conséquence la structure de données du dictionnaire, et écrivez une fonction pour indiquer si une phrase contient un syntagme correspondant à un élément du dictionnaire.

Question 5d

Disposant d'une phrase et d'un dictionnaire, donnez la meilleure traduction du point de vue de la pertinence.

Question 6 (4 points)

Bien, c'est maintenant que les choses sérieuses commencent ! Nous avons donc un système capable de renvoyer une traduction d'une phrase, et, abstraction faite des simplifications, nous devrions obtenir des traductions correctes [2]. Il reste maintenant à respecter le registre de langue.

On considère que le niveau de langue d'une phrase est la moyenne du niveau de langue des mots qui la composent, et on définit la fidélité d'une traduction comme le produit de la pertinence de cette traduction par la différence de registre entre les deux phrases.

Modifiez votre fonction de traduction afin qu'elle soit la plus fidèle possible.

Question 7 (2 points)

Quel est le temps d'exécution de votre algorithme pour un texte de 1000 mots avec un dictionnaire de 50000 entrées sur un iPhone, équipé d'un processeur ARM à 620 MHz ?

Un point supplémentaire sera accordé si la copie est correctement présentée.

2. Bon, en vrai, ce n'est probablement toujours pas le cas malheureusement...

Questions bonus

Ces questions peuvent vous rapporter des points seulement si vous avez répondu juste à toutes les questions précédentes.

Première question bonus

Même un texte grammaticalement correct peut être considéré comme incorrect. En français par exemple, on a tendance à vouloir éviter les répétitions et préférer utiliser des synonymes lorsque c'est possible. Cependant, ce peut aussi être un effet de style tout à fait volontaire.

Considérons qu'à partir de deux utilisations d'un même mot sans au moins une phrase entre les deux il y a répétition.

Modifiez votre fonction pour que la traduction d'un texte se fasse en respectant le style du texte original, à savoir utiliser des répétitions lorsqu'il y en a, et les éviter lorsqu'il n'y en a pas.

Deuxième question bonus

Traduisez cette phrase dans la langue de votre choix, représentez la sous forme de graphe syntaxique dans les deux langues, et montrez les transformations permettant d'aller de l'une à l'autre.

Troisième question bonus

Vous aurez sans doute remarqué que tout au long de ce sujet, nous avons considéré que nous avions déjà une représentation logique de la phrase à traduire. Autrement dit l'analyse lexicale était déjà faite. La raison en est que l'analyse morphosyntaxique des lexèmes d'une phrase (déterminer la nature des mots) est extrêmement complexe.

Puisqu'il vous reste du temps, indiquez en quoi cette analyse est complexe, et proposez quelques pistes. La solution complète à ce problème est également acceptée.