Polytech Paris Sud > Et4 Informatique > Recherche d'information dans les textes

Description du projet de groupe

Ce projet est à réaliser en groupe et sera à rendre collectivement. Le rendu sera composé des codes source développés dans le cadre du projet, ainsi que d'un rapport (moins de 20 pages) décrivant l'ensemble du projet et les contributions de chacun des intervenants. La date de rendu et de soutenance sera décidée ultérieurement dans la semaine du 5 mai 2014.

1 Vue d'ensemble du projet

L'apprentissage des langues étrangères est entré dans une nouvelle dimension avec l'essor d'Internet [1]. Il est désormais très facile de communiquer et de lire de très nombreux textes dans la langue d'apprentissage. Cependant, l'identification et le choix de textes appropriés est un problème difficile, qui dépend notamment du profil d'apprentissage de chaque personne.

En considérant ce problème de façon simplifié, on peut estimer qu'un bon texte support pour l'apprentissage d'une langue va employer à la fois des mots bien maîtrisés par l'apprenant, et des mots moins bien maîtrisés voire inconnus, mais en petite quantité (grosso modo, il s'agit donc de trouver des textes du niveau de lecture de l'apprenant, mais présentant tout de même une certaine nouveauté). En outre, il peut être intéressant de disposer d'une traduction de ce texte support dans une langue bien maîtrisée par l'apprenant, en particulier sa langue maternelle. C'est sur ce principe et sur la disponibilité de grandes quantités de textes en plusieurs langues que va reposer ce projet de recherche d'information dans les textes : il s'agira de trouver et proposer automatiquement des textes utiles à un apprenant particulier parmi une grande collection de textes.

2 Cahier des charges

Cette section décrit les différentes parties fonctionnelles du projet, dont certaines seront indiquées comme optionnelles.

2.1 Analyse et indexation des corpus

Le scénario principal sera celui d'une utilisation du système par un apprenant francophone de l'anglais. Les textes qui intéresseront donc tout particulièrement

1. Voir par exemple le site Lang-8 : `http://lang-8.com`

l'apprenant seront en anglais, et auront idéalement une phrase associée en français [2], par exemple [3] :

anglais *I think your theory does not hold water.*
français *Je crois que ta théorie ne tient pas debout.*

Les corpus pourront donc être indexés en plusieurs langues, et les phrases en relation de traduction devront pouvoir être mises facilement en correspondance. De plus, les phrases auront pu éventuellement subir des traitements appropriés pour faciliter leur usage. [4]

Dans ce projet, il n'est pas demandé à l'utilisateur de pouvoir ajouter son propre corpus. Les index seront donc construits préalablement pour permettre une utilisation efficace par l'utilisateur. Le corpus utilisé sera le corpus du projet communautaire `tatoeba` [5], qui met à disposition un grand ensemble de phrases décrites dans un certains nombre de langues, et donc notamment en anglais et en français pour un certain nombre d'entre elles.

2.2 Recherche d'exemples

La fonctionnalité la plus simple pour l'utilisateur sera la possibilité de chercher des extraits de textes (ici, des phrases issues de `tatoeba`), soit dans la langue d'apprentissage (l'anglais), soit dans sa langue maternelle (le français). En utilisant les index décrits section 2.1, il sera possible de trouver efficacement des phrases répondant "plus ou moins bien" à la requête posée. [6]

De plus, il devra être possible d'effectuer très efficacement des recherches d'expressions littérales telles que "*does not add up*", en utilisant un *tableau de suffixes*.

Optionnel À la manière du système `Linguee`, il peut être utile de mettre en évidence pour l'utilisateur les parties qui se correspondent entre deux phrases traduites. Pour faire cela, il faut connaître un *alignement* entre les mots de deux phrases en relation de traduction. L'outil `BerkeleyAligner` [7] permet par exemple de calculer des modèles d'alignement statistiques sur tout un corpus d'apprentissage (ici, l'ensemble des paires de phrases anglais-français du corpus `tatoeba`).

2. D'autres configurations de langues seront bien entendu les bienvenues.
3. cf. `http://tatoeba.org/fre/sentences/show/2397571`
4. Voir par exemple les outils de `OpenNLP` (`https://opennlp.apache.org`) ou de `StanfordNLP` (`http://nlp.stanford.edu/software`)
5. `http://tatoeba.org/fre/downloads`
6. Voir, par exemple, le résultat de la recherche en français "*ne tient pas debout*" en bilingue français-anglais avec le système `Linguee` :
`http://www.linguee.fr/francais-anglais/search?source=auto&query=ne+tient+pas+debout`
7. `https://code.google.com/p/berkeleyaligner`

2.3 Construction d'un graphe lexical de l'apprenant

Le système proposera des fonctionnalités de proposition automatique de textes d'apprentissage (section 2.4) et de textes d'évaluation (section 2.5). Pour cela, le système doit avoir une certaine connaissance des textes que maîtrise bien l'utilisateur. Ce dernier peut par exemple fournir un petit texte initial composé uniquement de mots bien connus. À partir de cela, le système maintiendra un *graphe lexical de l'apprenant* : les nœuds correspondront aux mots (sous une certaine forme), associés à un *niveau de maîtrise*, et les arcs porteront une *valeur de similarité* entre les mots de chaque nœud relié. [8]

Le niveau de maîtrise sera utilisé pour vérifier la bonne maîtrise des mots présents dans le graphe (cf. section 2.5) ; il pourra par exemple être initialisé à une valeur minimale, et être incrémenté à chaque fois qu'un exercice qui l'utilise est réussi.

La valeur de similarité devra rendre compte d'une certaine proximité sémantique entre mots. L'outil `word2vec` [9], fondé sur des calculs distributionnels sur corpus, pourra être utilisé pour construire des vecteurs pour chaque mot d'un corpus (par exemple, le corpus anglais de `tatoeba`) puis une mesure de similarité entre mots. [10]

2.4 Recherche de textes d'apprentissage

Étant donné un graphe lexical d'un apprenant et un corpus indexé, on souhaite trouver automatiquement des phrases qui serviront de bons exemples d'apprentissage pour l'utilisateur. Le principe général est que ces phrases devront contenir des mots nouveaux mais en nombre très limité. Par exemple, si l'apprenant connaît déjà bien les mots *he, hold, his, book, under, his*, la phrase *He is holding his books under his arms* peut constituer une phrase intéressante.

Un autre critère à prendre en compte sera la proximité d'un mot nouveau à apprendre (et donc à trouver dans un bon exemple d'apprentissage) avec les mots déjà présent dans le graphe lexical de l'apprenant. Par exemple, si l'apprenant connaît notamment bien déjà les mots *hand, body, elbow*, le mot *arm* peut être intéressant à apprendre.

Optionnel Il s'agirait donc dans le scénario précédent d'étendre le graphe de l'apprenant par enrichissement de "zones conceptuelles" déjà denses (par exemple, un domaine tel que le "corps humain", sur lequel l'apprenant a déjà de bonnes connaissances). À l'inverse, il peut être également intéressant de chercher à aider à apprendre des mots dans des domaines nouveaux. Pour cela, le système pourrait

8. Il sera possible de ne pas représenter un graphe complètement connecté : par exemple, les paires de mots ayant une similarité inférieure à un seuil choisi pourront ne pas être reliées par un arc.

9. `https://code.google.com/p/word2vec`

10. Pour le calcul de similarité utilisant les vecteurs construits par `word2vec`, une réimplémentation à partir du programme d'exemple (en langage C) sera nécessaire.

proposer la fonctionnalité suivante : l'utilisateur soumet au système un (court) texte dans la langue d'apprentissage qu'il aimerait pouvoir lire. Le système va donc l'aider à préparer cette lecture en identifiant quels nouveaux mots doivent être appris.

2.5 Recherche de textes d'évaluation

Optionnel La recherche de textes d'apprentissage (section 2.4) permet de trouver des phrases utiles pour apprendre de nouveaux mots. La recherche de textes d'évaluation permet ici de trouver automatiquement des phrases denses en mots du graphe lexical de l'apprenant, mais possiblement avec un faible niveau de maîtrise. Cela permet de constituter des jeux d'entraînement sur un vocabulaire supposé connu, mais imparfaitement. Par exemple, la phrase d'évaluation peut être présentée à l'utilisateur puis la phrase traduite dans sa langue maternelle peut être dévoilée dans un second temps. Seuls les mots bien compris verront leur valeur de niveau de maîtrise augmenter. Les mots bien maîtrisés ont vocation à ne plus être l'objet d'évaluations.

2.6 Système général

Le système général devra permettre *a minima* à l'utilisateur de parcourir efficacement le corpus utilisé par le système, ainsi que d'avoir accès à des propositions automatiques de textes d'apprentissage. Les aspects IHM pourront être travaillés afin de rendre compte de façon appropriée de l'intérêt relatif de chacun des résultats proposés. Les calculs devront être les plus efficaces possibles, aussi bien en mémoire qu'en calculs, afin de permettre une montée en charge du système.

Les grandes fonctionnalités accessibles à l'utilisateur sont :

1. recherche par mots clés ou fragments dans les documents de la langue d'apprentissage ou d'une autre langue disponible, et présentation des résultats par pertinence décroissante
2. affichage bilingue des résultats, avec mise en évidence des liens de traduction (*optionnel*)
3. consultation du graphe lexical de l'apprenant (*optionnnel*)
4. recherche de nouveaux textes d'apprentissage, qui pourront être trouvés par des critères automatiques (algorithme spécifique d'enrichissement du graphe lexical), par interaction avec le graphe lexical, ou par soumission d'un texte que l'utilisateur souhaite pouvoir lire
5. recherche de textes d'évaluation, interaction avec l'utilisateur pour vérifier quels mots ont été bien compris ou non et mise à jour des valeurs de *niveau de maîtrise* des mots dans le graphe lexical, et possibilité d'affichage de statistiques sur les mots posant problème à l'apprenant (*optionnel*)