Extraction et classification automatique de matériaux textuels pour la création de tests de langue

Murielle Marchand
Centre de traitement électronique des documents - CETEDOC
Université catholique de Louvain
Place Blaise Pascal, 1
1348 Louvain-la-Neuve
marchand@tedm.ucl.ac.be
Date de la thèse (à déterminer)

Mots-clefs – Keywords

Automates à états finis - analyse de corpus - extraction automatique de phrases - classification automatique de phrases - INTEX

Finite-state automaton - corpus analysis - automatic sentence extraction - automatic sentence classification - INTEX

Résumé – Abstract

Nous présentons l'état de développement d'un outil d'extraction et de classification automatique de phrases pour la création de tests de langue. Cet outil de TAL est conçu pour, dans un premier temps, localiser et extraire de larges corpus en ligne du matériel textuel (phrases) possédant des propriétés linguistiques bien spécifiques. Il permet, dans un deuxième temps, de classifier automatiquement ces phrases-candidates d'après le type d'erreurs qu'elles sont en mesure de contenir. Le développement de cet outil s'inscrit dans un contexte d'optimalisation du processus de production d'*items* pour les tests d'évaluation.

Pour répondre aux exigences croissantes de production, les industries de développement de tests de compétences doivent être capable de développer rapidement de grandes quantités de tests. De plus, pour des raisons de sécurité, les *items* doivent être continuellement remplacés, ce qui crée un besoin d'approvisionnement constant. Ces exigences de production et révision sont, pour ces organisations, coûteuses en temps et en personnel. Les bénéfices à retirer du développement et de l'implantation d'un outil capable d'automatiser la majeure partie du processus de production de ces *items* sont par conséquents considérables.

We present here the state of development of an automatic sentence extractor and classificator for use in the creation of language tests. This NLP tool has been designed to, in a first instance, automatically locate and extract from designated on-line databases candidate source sentences meeting specific linguistic criteria and, in a second instance, to classify those sentences according to the specific types of errors they are capable of supporting. The

development of this NLP tool is couched in the context and goals of automated item models instantiating for educational assessment.

To meet increasing testing demands, assessment industries must be able to quickly produce large number of tests and regularly replace the items to prevent lapses in test security. The high number of items that must be continuously 'retired' and replaced creates a great need for continuous item supply. Those high production and revision demands, apart from being time-consuming, are also costly. The development and implementation of a NLP tool capable of automating the bulk of the processing involved in instantiating the test item models is thus of considerable benefit to educational testing organisations.

1 Introduction

ETS (Educational Testing Service[1]) est une organisation américaine qui développe et administre chaque année plus de 12 millions de tests de compétences. Ces tests ont pour but d'évaluer les aptitudes des étudiants à divers moments de leur parcours scolaire et académique (le SAT, par exemple, doit être passé par les étudiants pour entrer à l'université ; parmi les autres tests les plus connus on trouve le GRE, le GMAT et le TOEFL[2]). Pour répondre aux exigences croissantes de production et de renouvellement de ses tests, ETS s'est tourné vers la recherche en traitement automatique du langage dans le but de créer des techniques permettant, d'une part, d'aider à la fabrication des tests, et d'autre part, d'automatiser les tâches de correction de tests. Dans ce domaine par exemple, ETS a développé *E-rater*, un système de TAL qui permet d'évaluer automatiquement la qualité d'un essai (Burstein, 1999 ; Powers, 2000).

Le PPST (*Pre-Professional Skills Tests*), l'un des nombreux programmes de tests développé par ETS, vise à évaluer au niveau national les compétences des futurs enseignants. Il se compose de trois parties destinées à évaluer, respectivement, les aptitudes de l' étudiant en mathématique, en lecture et en écriture. Le présent papier porte sur cette dernière partie du PPST et plus précisément sur des tests à choix multiples : les *multiple-choice writing items*. Ces tests ont pour but de mesurer les compétences linguistiques à l'écrit d'un examiné en termes d'accord pronom-antécédent, d'accord sujet-verbe, de coordination, de subordination, de constructions comparatives et d'autres conventions de l'anglais écrit[3].

Automatiser la création et le renouvellement des tests d'évaluation offre des bénéfices considérables tels que l'augmentation de la capacité de production, la réduction du coût de développement par *item* et à terme la réduction des exigences de pré-tests[4]. Cette automatisation devrait donc permettre de mieux répondre à certaines exigences liées à la sécurité des tests. En effet, le fait que les tests puissent être mémorisés et distribués à d'autres examinés dans un laps de temps relativement court limite le nombre d'utilisation que l'on

[1] http://www.ets.org
[2] http://www.ets.org/tests.html
[3] Pour des raisons de confidentialité, nous ne pouvons donner ici la liste exhaustive des critères linguistiques utilisés.
[4] Le phase de pré-test consiste en l'évaluation de la difficulté des tests crées avant la distribution officielle de ceux-ci. Un groupe de personnes passent les tests après leur création et l'on évalue la difficulté de chaque *item* d'après les résultats de ce groupe témoin.

peut en faire avant de les déclarer 'inutilisables' (*retired*). Ainsi, pour les tests diffusés à une échelle internationale, comme le GRE que des milliers d'examinés passent chaque année, l'exposition se doit d'être limitée dans le temps, et cela implique une production constante de nouveau matériel. Afin de répondre à cette demande de production à grande échelle, ETS emploie un grand nombre de personnes chargées de concevoir et valider les *items*. Ces processus de validation ne sont pas seulement coûteux, ils sont également peu rentables. En général, 1/3 seulement des tests produits passent avec succès la phase de révision et sont jugés aptes à l'utilisation.

Nous présentons ici l'état de développement d'un outil d'extraction et de classification automatique de phrases dont le but est d'automatiser la majeure partie du processus de création des *multiple-choice writing items* du PPST. Il s'agit plus précisément de l'élaboration d'un outil de TAL permettant, dans un premier temps, d'extraire automatiquement des phrases sur la base de critères linguistiques, numériques (nombre de mots) et thématiques ; et dans un deuxième temps, de classer ces phrases d'après les catégories citées plus haut (ex. construction comparative) et d'après le type d'erreur qu'elles sont en mesure de contenir, c'est-à-dire leurs sous-catégories (ex : construction comparative - sous-catégorie : « *less A than B* », « *more A than B* », etc.). Une prochaine phase du projet visera à intégrer au système un « modèle de difficulté » permettant de prédire la difficulté d'un *item* sans avoir à le pré-tester sur un groupe témoin. Ce « modèle de difficulté » se basera sur certaines caractéristiques linguistiques des phrases utilisées dans les tests[5].

1.1.1 Les PPST Multiple-choice writing items

Dans le PPST, les *items* sont des phrases d'une longueur d'environ 20 à 45 mots dans lesquelles 4 mots ou syntagmes sont soulignés. L'examiné doit choisir entre 5 options ; soit débusquer l'erreur et opter pour l'option A à D, soit juger que la phrase est correcte et opter pour l'option E '*no error*'. Dans la figure 1 ci-dessous, l'erreur se situe au niveau A et vise à évaluer l'utilisation correcte par l'examiné de la construction « *Between A and B* ».

Between 1947 to 1973, the average worker's earnings,
A
when adjusted for inflation, rose 61 percent,
B
reflecting a rise in productivity. No error
C D E

Figure 1 : Exemple de *item* pour le PPST

Les phrases-*items* dont est constitué le PPST sont généralement inspirées de phrases extraites de larges corpus, en général des articles de magazines ou des articles de journaux en ligne. Ces corpus doivent être continuellement renouvelés afin de répondre à la demande constante de production de phrases-*items*. Pour répondre à ce besoin continuel de nouveaux textes,

[5] ETS a développé une échelle de difficulté des items allant de -3.0 à +3.0. Par exemple, une erreur de majuscule dans une phrase est considérée comme très facile à détecter et est catégorisée -3.13 sur l'échelle de difficulté, tandis qu'une erreur de ponctuation (virgule) a une difficulté de -1.41. Ces estimations statistiques sont obtenues par l'analyse des résultats des étudiants dans les archives de plusieurs années de tests.

nous avons connecté notre système d'extraction à un système qui récupère continuellement de nouveaux textes sur internet (GlossaNet - Fairon, 1999).

2 Conception d'un outil d'extraction et de classification automatique de phrases

L'outil sur lequel porte le présent papier a pour but d'automatiser la sélection de phrases-*items* qui approvisionnent le PPST. Il est conçu à partir du système INTEX[6] pour localiser dans des bases de données en ligne (principalement des journaux) des phrases répondant à certains critères linguistiques et donc susceptibles de contenir certains types d'erreurs spécifiques. Le système extrait et classifie ces phrases d'après ces critères linguistiques pour ensuite les proposer comme phrases-candidates à un développeur de test qui y introduira manuellement l'erreur désirée (un projet ultérieur envisagera la génération automatique d'erreurs calibrées).

Une fois terminée la récupération de corpus par GlossaNet commence le processus d'extraction et de classification automatique des phrases. Celui-ci se déroule en 3 temps, comme illustré ci-dessous (étapes 2 à 4) :

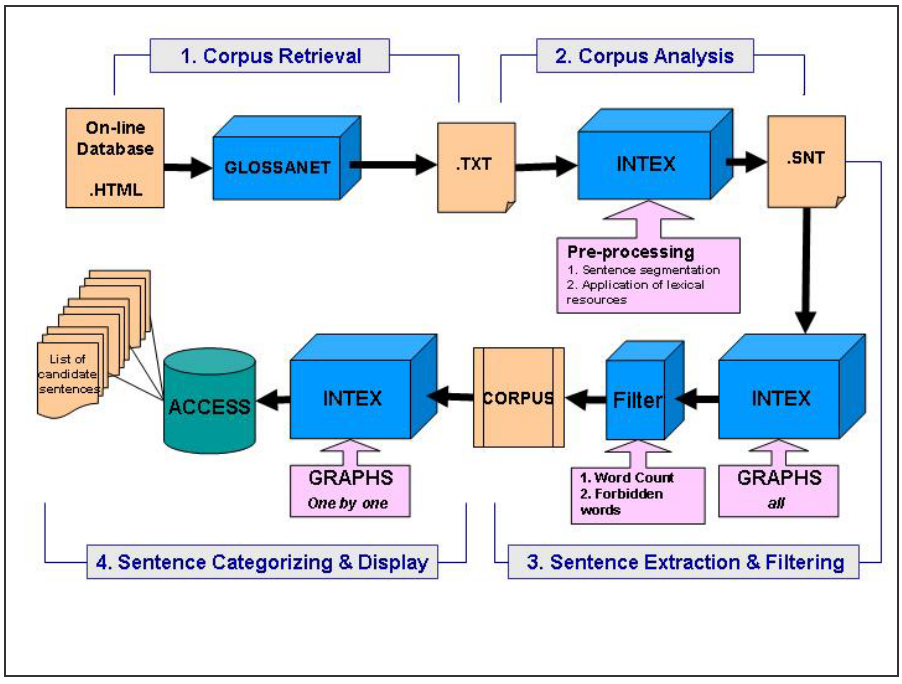

Figure 2 : Extraction et de classification de phrases à partir de corpus de données en ligne : méthodologie

L'analyse comprend :

[6] INTEX est un analyseur de corpus développé par Max Silberztein. Pour une présentation des différentes applications INTEX, voir Fairon 1999 et Silberztein 2000.

- une phase d'analyse de ce corpus de données textuelles (c'est-à-dire la normalisation du texte par le système INTEX et l'application de dictionnaires électroniques à large couverture) ;
- une phase d'extraction et de filtrage des phrases-candidates ;
- une phase de classification et d'affichage de ces phrases par catégories.

2.1 Récupération de matériel textuel en ligne

La méthodologie utilisée lors de la récupération de matériel textuel à partir de bases de données en ligne est inspirée du système GlossaNet (Fairon, 1999). Dans ce système, un module de téléchargement sophistiqué alimente continuellement le système avec de nouveaux textes prélevés sur internet. Ces textes sont rassemblés et filtrés pour pouvoir être ensuite utilisés par un logiciel de traitement de corpus.

2.2 L'analyse du corpus

Pour qu'un fichier-texte puisse être traité par le système INTEX, il doit passer d'abord par une phase de normalisation (*pre-processing*). Il s'agit ici d'une analyse linguistique qui segmente le texte en unités, les phrases, et applique ensuite sur le texte des dictionnaires de mots simples et de mots composés ainsi que des transducteurs lexicaux.

2.3 L'extraction de phrases avec le système INTEX

La phase d'extraction des phrases-candidates consiste en l'application d'automates à états finis (cf. Figure 3 ci-dessous) qui localisent et extraient du corpus normalisé des phrases répondant à certaines propriétés linguistiques.

2.3.1 Création et application de graphes :

Toutes les caractéristiques linguistiques recherchées sont décrites dans des graphes ou 'grammaires locales' (Gross, 1997). Par exemple, dans la phrase-item présentée plus haut (figure1), l'erreur introduite visait à évaluer la capacité de l'examiné à utiliser correctement la construction « *between A and B* ». Afin de rechercher dans une base de données textuelles d'autres phrases capables de supporter ce même genre d'erreur (*between A to B)*, nous avons décrit ce phénomène linguistique dans le graphe présenté à la Figure 3.

L'application de ce graphe sur le fichier-texte qui constitue notre corpus nous donne toutes les séquences de mots correspondant à « *between A and B* » sous forme de concordance (voir Figure 4). Chacune des phrases ainsi extraites constitue une phrase-candidate dans laquelle on pourra introduire une erreur de type « *between A to B »* ou *« between A with B »*.

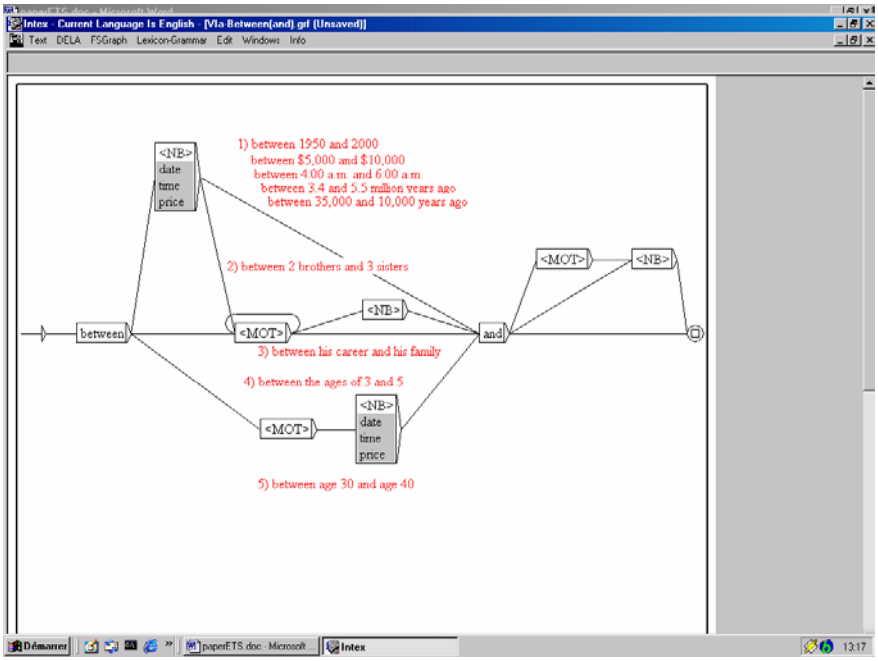

Figure 3 : Grammaire locale décrivant la séquence de mots « *between A and B* »

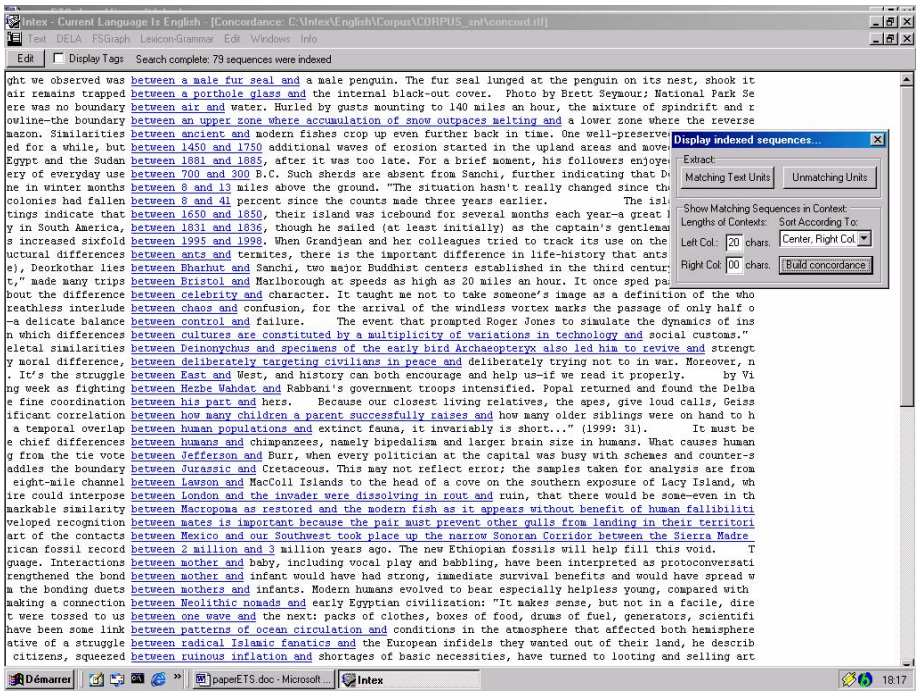

Figure 4 : Concordance de type « *between A and B* »

Les phrases dans lesquelles se retrouvent les séquences de mots recherchées sont extraites du corpus et sauvegardées dans un fichier indépendant. Ce même procédé de localisation et d'extraction est appliqué pour chacune des erreurs-types qu'ETS cherche à tester dans son programme PPST.

Il est important d'ajouter qu'à ce niveau-ci, nous appliquons sur le corpus de base tous les graphes simultanément. Cela afin de créer un nouveau corpus de matériel textuel qui ne

contienne que des phrases potentiellement porteuses d'une ou de plusieurs des propriétés linguistiques recherchées.

2.4 Le filtrage des phrases

L'étape suivante consiste en l'application sur ce nouveau corpus d'un filtre permettant de limiter encore le nombre de phrases-candidates sur base de critères numériques et thématiques cette fois-ci.

Le filtre numérique (qui traite de la longueur des phrases) dispose d'une option permettant au développeur de test de choisir le nombre de mots minimum et maximum que les phrases-candidates doivent contenir.

Le filtre thématique (développé en Perl) tient compte d'une liste de 'mots interdits' portant sur des sujets sensibles ou discriminateurs (en terme de sexe, culture, etc.). Ce filtre rejette également automatiquement les phrases contenant des anaphores, des conjonctions de subordination ou des pronoms faisant référence à un antécédent dans une phrase précédente. Les phrases doivent en effet être compréhensibles hors contexte.

Après le passage de ce double filtre, nous obtenons le corpus de phrases-candidates final, celui sur lequel nous allons maintenant pouvoir ré-appliquer chaque graphe séparément afin de classer chaque phrase par catégorie. C'est ce que nous expliquons au paragraphe suivant.

2.5 La classification des phrases

La phase de catégorisation et de classification des phrases consiste en l'application individuelle de chaque graphe sur tout le corpus. Après application d'un graphe, les phrases sélectionnées par celui-ci sont automatiquement étiquetées comme appartenant à la catégorie linguistique décrite par le graphe. Il va de soi qu'une phrase peut représenter plus d'une catégorie linguistique. Elle peut, par exemple, contenir une construction comparative, une coordination ainsi qu'un accord sujet-verbe intéressant pour le développeur de test. Ces données sont affichées et classées dans un fichier Access. En sortie, le développeur choisira la catégorie sur laquelle il veut travailler. Une liste de phrases appartenant à cette catégorie s'affichera avec, à côté de chaque phrase, le nombre de mots ainsi que les autres catégories linguistiques à laquelle appartient cette phrase. Le développeur pourra d'abord choisir de garder ou supprimer chaque phrase-candidate pour y introduire par la suite l'erreur désirée.

3 Conclusion

Nous avons présenté ici l'état de développement d'un outil d'extraction et de classification automatique de phrases dont le but est d'optimaliser la production de tests de langue pour ETS. Les aspects originaux du système développé sont, d'une part, la ré-utilisation et l'adaptation d'outils et de ressources génériques préexistants, et d'autre part, la création d'un système d'extraction autonome. De plus, nous avons expliqué en quoi un tel outil de TAL s'avérait maintenant indispensable pour les industries de production de tests d'évaluation. Un système d'extraction et de classification automatique de matériel textuel permet, en effet,

parce qu'il automatise la majeure partie du processus de construction de ces tests, un gain de temps et de personnel considérable.

Références

Bejar, I. I., Stabler, E. P., & Camp, R. (1987), Syntactic complexity and psychometric difficulty: A preliminary investigation, *ETS Research Report*, No. RR-87-25, Princeton, NJ, Educational Testing Service.

Burstein J., Chodorow M. (1999), Automated Essay Scoring for Nonnative English Speakers, *Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*.

DeMauro, G. E., Merritt, A., & Adams, R. (1994), Delimiting the verbal domain, *ETS Research Report*, No. RR-94-34, Princeton, NJ, Educational Testing Service.

Fairon C. (1999) ed., Analyse lexicale et syntaxique: Le système INTEX*, Lingvisticae Investigationes*, Tome XXII (volume spécial), Amsterdam/Philaldelphia, John Benjamins.

Fairon, C. (1999), Parsing a Web site as a corpus, in C. Fairon (ed.) Analyse lexicale et syntaxique: Le système INTEX*, Lingvisticae Investigationes* Tome XXII (volume spécial), Amsterdam/Philaldelphia, John Benjamins.

Gross M. (1997), The Construction of local grammars, in E. Roche et Y. Schabes (eds), *Finite State Language Processing*, The MIT Press, Cambridge MA.

Powers D.E., Burstein J.C., Fowles M.E., Kukich K. (2000), Comparing the validity of automated and human essay scoring, *GRE*, Vol. 98-08a.

Renouf A. (1988), Corpus Development, in J.M. Sinclair (ed.), *Looking up. An account of the COBUILD Project in lexical computing*, Londres-Glasgow, Collins ELT.

Silberztein M. (1993), *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Paris, Masson.

Silberztein M. (1998), Transducteurs pour le traitement automatique de textes, in B. Lamiroy (ed.), *Le Lexique-grammaire*, *Travaux de Linguistique 37*, Bruxelles, Duculot, pp. 127-142.

Silberztein M. (2000), *Manuel INTEX*. http://users.bestweb.net/~intex/downloads/Manuel.pdf

Sheehan, K. M. & Mislevy, R. J.(2001), An inquiry into the nature of the sentence-completion task: Implications for item generation, *ETS Research Report*, No. RR-01-13. Princeton, NJ, Educational Testing Service.