Convertisseur d'équations LATEX2Ink

A. Montaser Awal, R. Cousseau, C. Viard-Gaudin

Convertisseur d'équations LATEX2Ink

Ahmad Montaser Awal – Romain Cousseau – Christian Viard-Gaudin

IRCCyN – UMR CNRS 6597

Ecole polytechnique de l'université de Nantes

Rue Christian Pauc – BP 50609 – 44306 Nantes CEDEX 3 - FRANCE 1

{ahmad-montaser.awal, Romain.Cousseau, Christian.Viard-Gaudin}@univ-nantes.fr

Résumé *: Dans cet article nous présentons un outil de génération de formules mathématiques manuscrites en-ligne à partir d'une chaîne LATEX. Ce générateur permettra facilement de fabriquer à partir d'un corpus de référence d'expressions mathématiques une base de données qui sera annotée automatiquement au niveau symbole. Ainsi, à partir d'une base de symboles isolés, nous pouvons produire de façon pseudo-synthétique une formule mathématique quelconque par un placement et un dimensionnement stochastiques 2D de ces éléments. Nous montrons l'intérêt de cet outil dans le cadre d'un projet visant à la conception d'une méthode adaptée à la reconnaissance et à l'interprétation d'expressions mathématiques en-ligne.*

Mots-clés *:* Formule mathématique, reconnaissance structurelles, analyse syntaxique, génération d'expressions.

1 Introduction

Nous nous intéressons dans cette étude au problème posé par la reconnaissance des expressions mathématiques manuscrites en-lignes. Il s'agit là d'un vaste champ de recherche où peu de travaux internationaux existent, spécialement dans le domaine de l'écriture en-ligne. En particulier aucune base d'équations d'écriture en-ligne n'est aujourd'hui publiquement accessible. La disponibilité d'une telle base est pourtant une étape indispensable pour le développement, la mise au point et l'évaluation d'un système de reconnaissance d'expressions. De plus pour que cette base puisse être utilisée de manière efficace, celle-ci doit-être annotée. Il en résulte une tâche très fastidieuse pour réaliser cet étiquetage qui vient se rajouter à la collecte proprement dite de ces expressions. Comme alternative à cette approche, nous proposons de limiter la collecte aux seuls éléments terminaux du langage, que nous avons fixés ici à 223 symboles, et ensuite d'utiliser ces symboles pour alimenter un générateur d'expressions traduisant une chaîne LATEX quelconque en une expression manuscrite réaliste représentative de cette expression.

En effet, contrairement à un texte, les notations mathématiques utilisent un arrangement bidimensionnel de symboles pour encoder les informations. Être capable de comprendre et d'interpréter ce type d'expression implique de résoudre plusieurs problèmes [CHA 00]:

1. Segmentation,
2. Reconnaissance des symboles,
3. Interprétation de l'expression.

La phase de segmentation va consister à regrouper les points du tracé appartenant à un même symbole. A ce stade, il est nécessaire de préciser les hypothèses guidant cette étape de regroupement. La problématique de la reconnaissance de symboles est plus classique, encore qu'elle se trouve relativement plus complexe que pour la reconnaissance de caractères dans la mesure où le nombre de classes se trouve être relativement élevé. Enfin, l'étape d'interprétation va chercher à déduire de la disposition et de la taille relative de ces symboles une description structurée cohérente de l'expression.

2 Base de données

La création d'une base de données d'expressions est une étape indispensable pour permettre de concevoir, de développer et de tester un système de reconnaissance de formules mathématiques manuscrites en-ligne.

2.1 Base de données d'expressions

La première étape fût donc de rechercher un ensemble d'expressions mathématiques les plus diversifiées possibles couvrant les différents domaines utilisant des formules mathématiques. Pour cela, nous nous avons considéré les travaux effectués par Garain [GAR 05], qui était destiné à la reconnaissance de formules imprimées. Celui-ci est constitué d'expressions provenant de livres scientifiques d'universités écrits en anglais (la plupart de mathématiques). Ce corpus définit 100 expressions.

Ce corpus a été augmenté par celui provenant de la base Aster, tirée des travaux de Raman [RAM 94] qui regroupe elle aussi un ensemble d'équations (64 équations) intéressantes dans différents domaines.

2.2 Base de Symboles

L'ensemble des 223 symboles (chiffres, lettres grecques, flèche, …) les plus fréquemment utilisés, et notamment ceux permettant de reconstituer l'ensemble du corpus d'équations a été collecté à l'aide d'une technologie de type stylo/papier digital. Après récolte, la base de données est constituée de 62 440 échantillons de symboles isolés produits par 280 scripteurs.

2.3 Générateur d'équations

A l'inverse d'un système de reconnaissance, le générateur doit, à partir de la chaîne LATEX représentant l'expression, produire le tracé manuscrit et le sauvegarder sous le format Unipen. La création du fichier s'effectue en quatre étapes, FIG 1.

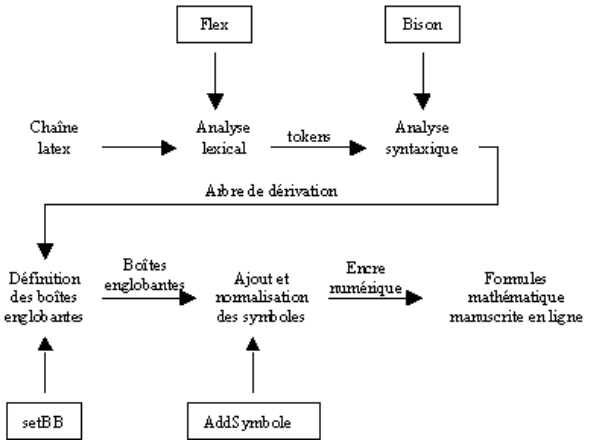

FIG. 1 – Processus de génération

La première des phases est l'analyse syntaxique de la chaîne LATEX. Elle est effectuée à l'aide de l'utilisation conjointe de deux outils : Flex et Bison.

Flex effectue l'analyse lexicale de la chaîne LATEX en entrée. L'analyse syntaxique effectuée par Bison, nous permet de construire l'arbre de dérivation de la chaîne analysée. La FIG. 2 montre l'exemple d'un arbre de dérivation obtenu à partir d'une chaîne LATEX.

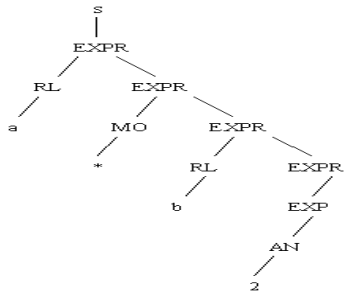

FIG. 2 – Arbre de dérivation pour la chaîne : $a * b^{2}$

Une fois l'arbre de dérivation de la chaîne LATEX obtenu, on effectue un parcours en « profondeur d'abord » afin de définir le placement et la taille des boîtes englobantes associées à chaque élément. Une fois que la position et la taille nominales de chaque symbole sont fixées, des perturbations aléatoires sont introduites, cela permet des réalisations différentes à partir d'un même ensemble de symboles pour une équation donnée.

3 Expérimentation et résultats

L'intérêt d'un tel outil vient du fait qu'il permet de développer la base de données d'expressions sans avoir recourir à de nouvelles phases de récoltes qui restent fastidieuses et consommatrices de temps.

Nous allons présenter quelques représentations obtenues à partir d'expressions extraites de notre corpus.

Il est difficile de mesurer la qualité objective de telles productions à défaut de disposer de critères d'évaluation. Nous sommes limités à visualiser les résultats obtenus et à les comparer avec les expressions correspondantes écrites globalement par un scripteur.

Ainsi, à la figure 3, (A) et (B) sont des résultats du générateur provenant de deux scripteurs différents. Par contre, (C) présente une expression manuscrite écrite globalement par un scripteur. D'un point de vue visuel, nous sommes capables de générer avec un rendu réaliste n'importe quelle expression désirée afin d'enrichir la base de données des expressions.

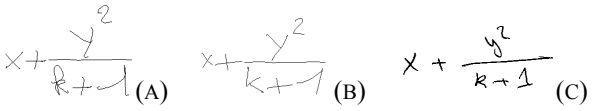

FIG. 3 –Expression simple

Néanmoins, le générateur réalisé ne vise pas à remplacer les expressions réelles, son but est de fournir une large quantité d'exemples afin d'être capable de tester des situations et des expressions quelconques. Cela ne dispensera pas à terme de tester le système de reconnaissance avec des expressions réelles.

4 Conclusion et perspective

Notre travail s'est situé en amont du processus de reconnaissance d'équations manuscrites en-ligne. Il a consisté en premier lieu, à une recherche d'un corpus d'équations mathématiques le plus général possible couvrant les différents domaines scientifiques. Nous avons développé un générateur d'équations synthétiques. L'application permet de générer l'encre numérique correspondant à l'équation mathématique décrite par sa chaîne LATEX en entrée. Cet outil présente l'avantage de permettre de tester facilement le comportement d'un système donné de reconnaissance vis-à-vis d'une équation quelconque. Par contre, il ne prétend pas mimer de façon exacte le comportement d'un scripteur humain.

Références

[CHA 00] K.-F. Chan and D.-Y. Yeung. Mathematical Expression Recognition : A Survey. International Journal on Document Analysis and Recognition, Volume 3, No.1, 2000, pp. 3-15.

[GAR 05] U. Garain. Automatic recognition of printed and handwritten mathematical expressions. Thesis, Indian statistical institute, 2005.

[LEH 96] S. Lehmberg, H-J. Winkler, and M. Lang. A soft-decision approach for symbol segmentation within handwritten mathematical expressions. In ICASSP'96, pages 3434-3437.

[MAR 71] W. Martin, Computer input/output of mathematical expressions. Proc. 2nd Symp. on Symbolic and Algebraic Manipulations, New York, 1971, 78-87.

[RAM 94] T. V. Raman. Audio system for technical reading, thesis, Cornell university, 1994.