DEVELOPPEMENT D'OUTILS D'ANALYSE STATISTIQUE TEXTUELLE

A. Pibarot, L. Denoue, D. Labbé et J. Picard

(Publié dans : *Travaux scientifique du Service de Santé des Armées*. XVI, 1995, p. 305-307.)

Dans le domaine de la santé, du moral et des conduites à risque les questionnaires et les interviews restent un moyen d'approche irremplaçable. Ce domaine intéresse au plus haut point le commandement pour qui le facteur humain, c'est-à-dire l'homme dans sa diversité et sa variabilité, reste la composante essentielle d'une bonne défense et ce d'autant plus que les matériels et les missions se complexifient. Toutefois les enquêtes offrent une grande quantité d'informations souvent inexploitées ou mal exploitées par manque d'outils adaptés.

C'est pour répondre à ce besoin que le BMI développe des logiciels d'analyse textuelle susceptibles d'extraire les variables significatives et si possible les principaux thèmes abordés dans des textes libres. L'un des principaux problèmes soulevés par l'analyse automatique est la place importante des formes homographes (près de 30% dans un texte courant) que seules des analyses de type grammatical ou syntaxique pourront s'efforcer de lever.

Dans un premier temps, cette opération a été réalisée par un logiciel développé sous Macintosch au laboratoire du CERAT de l'Université des sciences sociales de Grenoble et porté au CRSSA sur des plates-formes DOS/Windows (1). Toutefois, les programmes et les tables de mots ou expressions avaient été conçues pour des machines de faible capacité, ce qui allongeait le temps de traitement, tout en imposant une décomposition des étapes et d'assez nombreuses interrogations. C'est pourquoi, en collaboration avec le CERAT, l'ensemble du système de traitement a été revu, en gardant l'optique d'une utilisation par des machines de bureau.

Nous présenterons successivement la configuration actuelle du logiciel et une application à la levée d'ambiguïtés. Suivra une brève discussion sur les problèmes posés par les analyses automatiques.

CONFIGURATION DU LOGICIEL

Il s'agit d'un logiciel de lemmatisation qui fournit, pour chaque mot du texte, une forme canonique (l'entrée dans le dictionnaire) et un code grammatical avec le genre pour les substantifs et, pour les verbes, quelques indications sur leur conjugaison (fléchie, participe présent, participe passé, infinitif). L'analyse de type sémantique, utilisée pour la détection de thèmes, fait l'objet d'une autre étape et de logiciels spécifiques qui ne sont pas présentés dans cette note.

Le nouveau logiciel de lemmatisation présente trois caractéristiques principales :

1 Fusion des étapes de mise à la norme et de codification initiale

La normalisation doit permettre de reconnaître le même lemme sous des variantes (ex. auto-stop = autostop) et surtout de ne pas séparer des mots ou expressions qui sont traités comme une unité dans le dictionnaire (de bric et de broc). Le programme qui

auparavant se référait aux tables pour reconnaître ces expressions et les coder unifie désormais ces deux opérations, ce qui représente un gain de temps important

Concrètement, il dissocie systématiquement tous les mots d'un texte et réagglutine (c'est-à-dire met sur une seule ligne, correspondant à une entrée) ceux qui correspondent à certaines règles : en particulier quand ils figurent dans les tables de mots composés sous une forme complète ou comme développement possible d'un préfixe (franco-, anti-...). Par contre certains mots, même reliés, seront logiquement dissociés, comme les pronoms (dit-il) ou le t euphonique (dira-t-on), sauf à sauvegarder quelques expressions même ambiguës (rendez-vous).

Les mêmes règles s'appliquent aux mots à majuscules. Et les formes n'obéissant pas aux critères retenus sont systématiquement dissociées, ce qui permet de lutter contre les liaisons abusives (économique-et-social).

Les mêmes règles s'appliquent aux mots à majuscules dont l'emploi est également soumis à vérification. Un mot à majuscule non reconnu comme un nom propre sera traité comme un nom commun, tout en gardant trace de la forme originale dans la case mémoire associée au mot. S'il n'est pas trouvé dans les noms communs, plusieurs options sont possibles selon le paramétrage du programme (cf § 3)

2 Standardisation de la codification

Avant l'interrogation éventuelle de l'opérateur, le nouveau programme comprend deux étapes de traitement automatique, auxquelles correspondent deux types de codification :

- Le traitement initial, décrit au § 1, aboutit à un fichier dont chaque ligne contient un mot, simple ou composé, avec en regard son lemme et son code grammatical (ou les lemmes et codes éventuels en cas d'homographie).
- Le traitement final s'appuie sur l'analyse contextuelle pour la levée des ambiguïtés et retourne à l'opérateur les cas non résolus en offrant plusieurs solutions précodées.

Dans la version actuelle, les codes sont standardisés à 4 chiffres pour le mot et 3 pour l'indice d'homographie. Le deuxième chiffre des codes de mot est un indice de construction désormais généralisé à tous les mots (même s'il est égal à zéro), ce qui peut s'avérer très utile lors de la phase de levée des ambiguïtés par analyse de contexte.

Si le mot n'a pas été identifié automatiquement, la procédure dépendra de l'option choisie lors du paramétrage.

3 Choix entre plusieurs options de traitement

Les sources potentielles d'erreur (ambiguïtés, variantes, fautes d'orthographes) sont si nombreuses qu'il est difficile d'atteindre le zéro faute dans le traitement purement automatique de la langue. C'est pourquoi l'interrogation de l'opérateur est prévue, en fin de parcours, pour les mots non trouvés.

A côté de cette option, dite "traitement dur", une autre option, dite "traitement mou" permet de sauter la phase d'interrogation qui peut être longue dans un texte comportant beaucoup de noms propres non codés. Le mot à majuscule absent des tables est systématiquement conservé dans sa graphie initiale et les autres mots non trouvés sont codés "mots inconnus".

UNE APPLICATION

L'analyse contextuelle est un programme complexe, encore en voie d'expérimentation, en raison des multiplicité des homographies qui concernent généralement des mots fréquents.

En voici en exemple extrême : le mot tout et ses flexions, véritable "bonne à tout faire" de la langue française. Le problème est résumé dans le tableau suivant :

	déterminant	pronom	adverbe	nom
Tout	x	x	x	x
Toute	x	x	x	
Toutes	x	x	x	
Tout	x	x		

Le programme repose sur l'application d'un automate, c'est-à-dire d'un nombre fini de règles permettant de résoudre le maximum de cas. Les 12 cas présentés se ramènent pratiquement aux 4 règles suivantes :

1) Tout est déterminant (adjectif indéfini) quand il est employé dans un groupe nominal et qu'il est accordé aux autres éléments du groupe (tout le monde, tous deux).

2) Tout est pronom lorsqu'il est employé seul ou associé à un groupe verbal (il a tout su).

3) Tout est adverbe lorsqu'il est placé devant un adjectif ou employé dans une locution adverbiale ou prépositive (il est tout seul).

4) Tout est substantif quand il est précédé d'un déterminant ou d'une préposition et suivi d'autre chose que d'un substantif ou d'un adjectif (le tout pour le tout).

L'algorithme bute sur des cas impossibles à résoudre parce que dépendant de l'interprétation : "elles sont toutes contrites" = "toutes (pronom) sont contrites" ou "elles sont extrêmement (adverbe) contrites". De même pour "nous avons tous nos défauts".

En l'état actuel du programme, environ 10% des flexions (tous, toute, toutes) restent non codées du fait des télescopages possibles entre pronom, déterminant et adverbe. Certains homographies peuvent toutefois être résolues par la prise en compte de locutions (tout à coup, tout de même, après tout, en tout...).

DISCUSSION

"Tout" représente un cas extrême, dans la mesure où par exemple le déterminant peut ne pas s'accorder avec le nom qu'il détermine (j'ai lu tout Les Plaideurs), alors que l'adverbe, théoriquement invariable, pourra s'accorder pour des raisons euphoniques (elle est toute heureuse). Même si l'on récuse "l'absurdité de notre orthographe" (Valéry), il est irréaliste de vouloir résoudre automatiquement et sans erreur tous les cas d'homographie. Cela tient à la nature de la langue, système indéfini, ouvert sur le monde et en constant changement.

La programmation d'une analyse automatique impose donc des choix qui concernent à la fois l'informatique et le langage, ou du moins l'idée qu'on s'en fait.

Contraintes de l'informatique de bureau

La programmation amène à définir un nombre limité de règles et à minimiser l'incidence des cas particuliers, surtout lorsqu'on développe un logiciel pour micro-ordinateur multitâches. La prise en compte des substantifs composés et des locutions permet certes de lever quantité d'ambiguïtés, mais vouloir tout régler par des tables est une gageure car le nombre des compositions est indéfini. Comme il n'existe pas de critère strict pour distinguer les unités lexicales graphiquement complexes des syntagmes libres (2), la solution la plus sage est de se limiter aux entrées des dictionnaires. Par contre les variantes d'écriture peuvent être recherchées par programme lorsque l'usage est erratique (lèche vitrine = lèche-vitrine = lèchevitrine).

L'application des règles morphosyntaxiques

La sagesse conseille également de s'en tenir aux catégories grammaticales et d'appliquer de façon restrictive les règles morpho-syntaxiques traditionnelles, déjà assez complexes. En voici un exemple pour l'accord du participe passé :"ces pommes, vous les avez prises ?", mais "des pommes, vous en avez pris ?" ; sans oublier l'homographie avec la forme fléchie (tu prises les pommes ?) ou le substantif (prises de courant). Il n'est déjà pas facile, et pas toujours possible, de régler par programme la séparation entre participe passé et adjectif, entre adjectif et substantif...On peut raffiner à l'infini l'analyse, mais la complexité croît exponentiellement pour des gains d'information de plus en plus faibles. C'est pourquoi nous renonçons par exemple à tenir compte de la polyvalence fonctionnelle des numéraux qui peuvent faire fonction de déterminant, adjectif, substantif (le trois de pique) ou nombre.

Ce ne sont pas seulement les besoins de l'analyse automatique qui poussent à l'harmonisation de l'orthographe et, plus fondamentalement à la simplification grammaticale. Le recul actuel de la langue française dans le monde et les risques pour beaucoup de français de devenir étrangers à leur propre langue sont également matière à réflexion.

Références

1 J. PICARD, A. PIBAROT et D. LABBE - Un outil de statistique textuelle : le lemmatiseur. *S.S.A 1995 Trav. scient. n° 16*, 305-306.

2 J. PICOCHE - *Précis de lexicologie française*. Paris, Nathan, 1992.