Écriture automatique

Laurence Danlos

les cahiers de
L'INRIA
Rendez-vous avec les sciences et technologies de l'information et de la communication

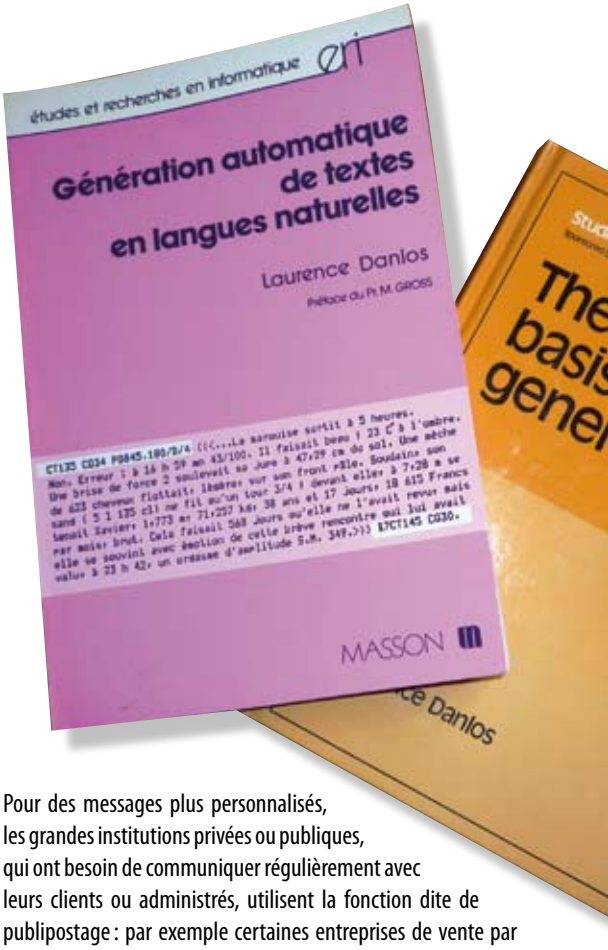
INFORMATIQUE-LINGUISTIQUE

Écriture automatique

Dans un grand nombre de domaines, la génération automatique de textes par ordinateur pourrait libérer l'être humain de tâches de rédaction fastidieuses et répétitives.

Il y a quelques mois, des chercheurs américains ont défrayé la chronique en annonçant être capables de faire écrire des articles journalistiques par un ordinateur. Même émotion qu'il y a vingt-cinq ans (fig. 1) : les journalistes pourraient-ils désormais être remplacés par des robots ? On en est évidemment très loin. La génération automatique de textes peut sans doute rendre un fier service dans certains domaines, mais pas nécessairement dans ceux que l'on imagine de prime abord. C'est ainsi que dans le cadre d'une collaboration avec Kantar Media*, nous avons pour notre part mis au point un système de production automatique de textes, EasyText*, destinés à commenter des tableaux de données statistiques. Mis en œuvre depuis cette année, c'est le premier système de cette nature à être opérationnel en France.

Depuis qu'ils existent ou presque, les ordinateurs produisent des textes en langue naturelle. Exemple : *« Votre imprimante n'a plus de papier »*. Mais ce type de message, qui s'affiche en cas de besoin après une commande d'impression, est préenregistré et ne nécessite donc aucune « intelligence ». Autres exemples, les messages en provenance d'administrations ou d'institutions telles que celui-ci, envoyé par une banque : *« Votre compte est débiteur de 144 €. Nous vous prions de l'approvisionner au plus tôt. »*. Dans ce cas, le message est produit automatiquement en complétant une « phrase à trou » : cette phrase est elle aussi préenregistrée et demande juste à être complétée par la valeur idoine, laquelle est extraite d'une base de données.



Pour des messages plus personnalisés, les grandes institutions privées ou publiques, qui ont besoin de communiquer régulièrement avec leurs clients ou administrés, utilisent la fonction dite de publipostage : par exemple certaines entreprises de vente par correspondance qui doivent répondre aux milliers de lettres de réclamation qu'elles reçoivent chaque jour. Cette fonction est notamment accessible dans des logiciels de traitement de texte tels que Word. L'opération de publipostage consiste à fusionner un tableau Excel et une lettre-type : les lignes du tableau correspondent aux différents clients, les colonnes enregistrent des

informations sur ces clients, la lettre-type contient des variables et des formules logiques (voir l'encadré « Une lettre-type pour publipostage »).

Le publipostage est certes très utilisé de nos jours, mais ses limites sont évidentes. D'abord, lorsque les données par individu sont nombreuses, c'est-à-dire lorsque le tableau Excel contient un grand nombre de colonnes, il devient très difficile de concevoir une lettre-type. Ensuite, le texte généré reste relativement élémentaire : il serait par exemple impossible de produire un commentaire qui ressemble à une synthèse « intelligente » et concise d'un tableau de chiffres, c'est-à-dire à autre chose qu'un texte qui soit la concaténation de paragraphes dont chacun ne ferait que paraphraser une ligne du tableau. C'est précisément à cette tâche que nous nous sommes attelés. Jusque-là, TNS media intelligence envoyait à ses clients des tableaux de chiffres dépourvus de tout commentaire. Notre objectif était de générer automatiquement un petit commentaire par tableau.

Comment fonctionne un système de génération automatique de textes ? Son entrée est un ensemble de données dont la forme dépend évidemment de l'application visée : quelques systèmes existent à ce jour dans les domaines de la biologie et de la médecine (résultats d'analyses physiologiques, compte-rendus médicaux…), pour la description d'itinéraire, la météo, la Bourse…, mais ils sont essentiellement en anglais. Dans le cas qui nous intéresse, le système a pour entrée un tableau de chiffres comme celui présenté page suivante (fig. 2), qui porte sur les investissements dans la téléphonie mobile. Il s'appuie sur un formalisme (G-TAG) capable de transformer une représentation conceptuelle d'un ensemble d'informations en un texte[1,2,3].

La première opération à effectuer consiste à sélectionner les informations significatives pour le client, autrement dit à répondre à la question « quoi dire ? ». Il s'agit donc d'extraire du tableau de chiffres les éléments saillants, par exemple les progressions des investissements entre 2008 et 2009 dépassant un seuil fixé. Cette opération, déterminante, relève plutôt du domaine de l'intelligence artificielle (système expert). Elle débouche sur un « message » représenté généralement sous forme logique.

Ensuite on doit passer de cette forme logique à un texte rédigé dans la langue de l'utilisateur (ou à plusieurs textes, rédigés en différentes langues). La question est alors : « comment le dire ? ». Ce passage à la langue naturelle s'effectue en deux étapes : une étape dite de macro-planification et une étape de micro-planification. La macro-planification consiste à « linéariser » la forme logique : celle-ci se présente souvent comme une conjonction (non ordonnée) de nombreuses formules logiques qu'il s'agit de transformer en une séquence cohérente de phrases. Pour être compris, un texte doit en effet

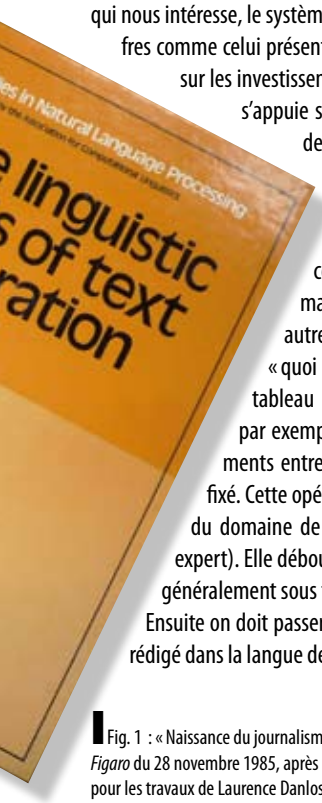


Fig. 1 : « Naissance du journalisme artificiel », tel était le titre d'un article du *Figaro* du 28 novembre 1985, après la remise du prix scientifique d'IBM-France pour les travaux de Laurence Danlos sur la génération automatique de textes. Dans ces travaux, publiés en français et en anglais, elle décrivait entre autres le système qu'elle avait mis au point pour faire rédiger par la machine (en anglais et en français) de courts récits d'attentats dans un style journalistique, les informations étant obtenues à partir d'un formulaire. L'article du *Figaro* indiquait qu'« il était désormais possible de réaliser des articles de journaux de façon automatique »… C'était oublier que l'ordinateur ne fait que ce qu'on lui dit de faire, en particulier de la compilation de données plus que de l'analyse.

Une lettre-type pour publipostage

La clientèle d'une entreprise peut être représentée sous forme d'un tableau dont chaque ligne correspond à un client et chaque colonne aux informations sur ces clients (sexe, nom, prénom, solde…). Le publipostage consiste à fusionner ce tableau et une lettre-type. L'exemple qui suit, où le symbole « $ » désigne le champ à remplir pour chaque client, montre une lettre-type censée avertir le client s'il est débiteur ou pas, avec le solde dû ou bien le crédit restant sur son compte. Les informations sont extraites d'une base de données actualisée :

{IF $SEXE=F Chère Cher} $PRENON $NOM,
Nous avons le {IF $SOLDE<0 regret plaisir} de vous informer que votre compte est {IF $SOLDE<0 débiteur créditeur} de $SOLDE…

De la phrase au texte

Le principal objet d'étude de la linguistique est la phrase. De ce fait, les linguistes ont développé des lexiques et des grammaires qui décrivent les contraintes observées au sein du domaine phrastique. Cependant, d'autres contraintes surgissent dès lors que l'on passe à l'enchaînement de plusieurs phrases en un texte cohérent.

Ainsi considérons la distribution d'un adverbe temporel (à midi) dans un discours causal composé de deux phrases, l'une décrivant la cause notée Pc, l'autre le résultat noté Pr. Il apparaît qu'une information temporelle peut figurer dans l'une ou l'autre phrase si la phrase décrivant le résultat précède celle décrivant la cause, soit Pr < Pc :

« Luc a fêlé la carafe à midi. Il l'a cognée contre l'évier. »
« Luc a fêlé la carafe. Il l'a cognée contre l'évier à midi. »

En revanche, elle ne peut figurer que dans la phrase de la cause si la phrase décrivant la cause précède celle décrivant le résultat, soit Pc < Pr :

« Luc a cogné la carafe contre l'évier à midi. Il l'a fêlée. »
« Luc a cogné la carafe contre l'évier. Il l'a fêlée à midi. »

En effet, le dernier énoncé a perdu toute interprétation causale (et devient même difficilement interprétable). Il faut donc établir des règles pour éviter de l'engendrer, ces règles sortant du domaine phrastique.

* La marque Kantar Media, fruit du regroupement de plusieurs sociétés d'informations marketing, dont TNS Media Intelligence, a été créée en janvier 2010. TNS Media Intelligence est l'une des branches du groupe TNS, qui comprend également TNS Sofres.

* EasyText a été développé par Frédéric Meunier, fondateur et gérant de la société Watch System Assistance, et Laurence Danlos, responsable scientifique de l'UMR ALPAGE (Inria-Rocquencourt, université Paris-Diderot).

* Sur ce sujet, voir par exemple : L'ère des robots-journalistes, dans *Le Monde* du 9 mars 2010.

INRIA

EVOLUTION DES INVESTISSEMENTS PAR SECTEUR / VARIETE										
ANALYSE	Dans le secteur TELEPHONIE MOBILE, les investissements enregistrent une baisse respective de 22% et 31% pour les variétés OFFRES PREPAYEES et OFFRES ABONN.GRAND PUBLIC en mars 2009 par rapport à mars 2008. Par ailleurs, depuis le début de l'année, pour la variété OFFRES ABONN.GRAND PUBLIC, ils enregistrent une baisse de 10%. Toutefois, sur la même période, la variété OFFRES PREPAYEES a triplé ses investissements (+198%).									
	Mars 2008	SOV	Mars 2009	SOV	Evol %	Cumul Janvier à Mars 2008	SOV	Cumul Janvier à Mars 2009	SOV	Evol %
TELEPHONIE MOBILE	60 747	100,0%	64 893	100,0%	7%	107 341	100,0%	131 134	100,0%	22%
OFFRES GRAND PUBLIC	37 819	62,3%	26 266	40,5%	-31%	54 231	50,5%	48 604	37,1%	-10%
OFFRES PREPAYEES	1 952	3,2%	1 519	2,3%	-22%	4 979	4,6%	14 857	11,3%	198%
Internet	3 331	5,5%	9 775	15,1%	193%	8 570	8,0%	14 575	11,1%	70%
OFFRES PRO	34	0,1%	5 266	8,1%	15 189%	88	0,1%	11 286	8,6%	12 768%
MARQUE	11 190	18,4%	4 521	7,0%	-60%	20 685	19,3%	10 602	8,1%	-49%
FORFAITS BLOQUES	745	1,2%	6 592	10,2%	785%	995	0,9%	9 258	7,1%	831%
OFFRES ENTREPRISE	931	1,5%	8 638	13,3%	828%	5 678	5,3%	8 907	6,8%	57%
INSTITUTIONNEL	2 467	4,1%	1 212	1,9%	-51%	9 320	8,7%	7 455	5,7%	-20%
OFFRES SERVICE DE CONTENU	2 277	3,7%	1 078	1,7%	-53%	2 440	2,3%	5 041	3,8%	107%
GAMME			26	.		355	0,3%	549	0,4%	55%

Fig. 2 : Ce tableau de chiffres visualise l'évolution des investissements en téléphonie mobile selon les secteurs entre mars 2008 et mars 2009. L'analyse qui en est faite est générée par EasyText de manière totalement automatique, sans révision humaine.

être bien structuré : il faut que les idées s'enchaînent les unes avec les autres de manière rationnelle et que l'ensemble du texte suive une progression thématique. La modélisation de ces principes s'appuie sur des connaissances rhétoriques et pragmatiques qui ont fait l'objet de divers travaux théoriques au cours des dernières années[4,5,6].

La macro-planification peut également s'appuyer sur des « schémas de textes » établis en étudiant un corpus de textes écrits manuellement et visant à communiquer le même type de messages que le système de génération. Enfin, cette opération se fonde sur des connaissances linguistiques qui émergent peu à peu au fur et à mesure que les efforts de recherche sortent du cadre de la phrase pour aborder celui du texte (voir l'encadré « De la phrase au texte »). Au final, la macro-planification se concrétise par une suite de « schémas de phrase » reliés par des relations de discours. Ces dernières sont de différentes natures : explication, résultat, narration, contraste, *etc*. Elles matérialisent le lien entre deux (groupes de) phrases successives. Ainsi dans l'énoncé « *Luc a quitté le XVI^e^ arrondissement. Son loyer était trop cher* », les deux phrases sont reliées par une relation de discours de type « explication », alors que dans l'énoncé « *Luc est parti vivre à Barbès. Il paie donc moins cher de loyer* », la relation de discours est de type « résultat ».

Quant à l'étape de micro-planification, elle consiste d'une part à déterminer, pour chaque relation de discours, s'il est possible et nécessaire de la préciser par un connecteur de discours (on parle de lexicalisation), qu'il soit de type adverbial ou en forme de conjonction : une relation de « résultat » pourra ainsi être précisée par la forme adverbiale « de ce fait », une relation d'« explication » par la conjonction « parce que », une relation de « narration » par l'adverbe « ensuite », *etc*. Pour chaque schéma de phrase, il faut d'autre part choisir les items lexicaux qui la composent, les constructions syntaxiques des verbes, les déterminants préfixant les noms, les positions respectives des compléments du verbe et des compléments circonstanciels, *etc*. L'objectif est en effet d'aboutir à une phrase bien formée de la langue cible.

Cette dernière opération repose sur des connaissances approfondies en sémantique, en syntaxe et en morphologie, même si on peut se contenter de générer des textes qui n'offrent pas toute la richesse de la langue. De plus, l'algorithme est complexe car les décisions ne sont pas indépendantes d'une phrase à l'autre. Il est en particulier nécessaire de respecter les contraintes de parallélisme entre phrases : par exemple, des phrases de même sujet (sujets coréférents) doivent de préférence se succéder. Au final, le texte produit ressemble à s'y méprendre à un texte produit par un humain (fig. 2).

Ce système de génération de textes permet ainsi de produire des commentaires sur des centaines de tableaux de chiffres envoyés chaque mois aux clients de Kantar Media. Il ne remplace toutefois pas une quelconque activité humaine : auparavant, aucun commentaire n'accompagnait ces mêmes tableaux car leur rédaction, au-delà de leur caractère fastidieux, aurait demandé une quantité considérable de personnel. Mais l'ordinateur n'invente pas ni n'analyse : il ne fait que de la compilation d'informations selon des règles qu'on lui donne. Cette remarque vaut pour le système américain Stats Monkey, mis au point par des chercheurs de NorthWestern University et relatant dans un style journalistique les faits marquants d'un match de baseball*. Pour notre part, nous avions déjà produit automatiquement des récits d'attentats (en français et en anglais) il y a vingt-cinq ans (fig. 1). Aucun journaliste ne s'est trouvé au chômage depuis du seul fait d'une telle pratique.

Laurence Danlos mathématicienne de formation, s'est spécialisée dans le traitement automatique des langues, notamment en génération automatique, traduction automatique et compréhension de textes. Elle dirige l'équipe ALPAGE, unité mixte de recherche (UMR) Inria-Rocquencourt et université Paris-Diderot.

(1) L. Danlos, G-TAG : un formalisme lexicalisé pour la génération de textes inspiré de TAG, *Revue T.A.L., vol. 39 (2)*, pp. 7-33, 1998

(2) F. Meunier, *Implémentation du formalisme G-TAG*, thèse d'informatique de l'université Paris-VII, 1997

(3) L. Danlos et F. Meunier, FLAUBERT : an user friendly system for multilingual text generation, *in Proceedings of the 9 th workshop on Natural Langugage Generation*, Niagara, pp. 44-56, 1998

(4) W. Mann et S. Thompson, Rhetorical structure theory : Toward a functional theory of text organization, Text, 8 (3), pp. 243-281, 1988

(5) N. Asher, *Reference to Abstract Objects in Discourse*, Kluwer, Dordrecht, 1993

(6) N. Asher et A. Lascarides, *Logics of Conversation*, Cambridge University Press, 2003