TALN 2003, Batz-sur-Mer, 11–14 juin 2003

Structuration automatique de preuves mathématiques : de la logique à la rhétorique

Adil El Ghali — LATTICE – PPS
Laurent Roussarie — LATTICE – CNRS
Université Paris 7 – Case 7003
2, place jussieu
75251 Paris Cedex 05
{adil,laurent}@linguist.jussieu.fr

Mots-clefs – Keywords

Génération automatique de textes, détermination de contenu, logiques de description, structuration de document, SDRT

NLG, content determination, description logics, document structuring, SDRT

Résumé

Nous présentons dans ses grandes lignes un modèle de structuration de documents pour la génération automatique de preuves mathématiques. Le modèle prend en entrée des sorties d'un prouveur automatique et vise à produire des textes dont le style s'approche le plus possible des démonstrations rédigées par des humains. Cela implique la mise au point d'une stratégie de planification de document capable de s'écarter de la structure purement logique de la preuve. La solution que nous proposons consiste à intégrer de manière simple des informations de type intentionnel afin d'enrichir la structure rhétorique finale du texte.

1 Introduction

En génération automatique, les stratégies de *structuration de document*[1] ont non seulement un impact durable sur l'ensemble du processus de production, mais aussi une efficacité double : elles constituent une première étape de *mise en forme* (les choix des structures de discours se répercutent sur l'agencement final du texte), et elles incluent une phase de *raisonnement* qui n'est pas complètement indépendante de la détermination de contenu : il s'agit, par exemple,

[1] Pour des détails sur l'architecture et les composants des systèmes de génération, voir (Reiter & Dale, 2000).

de construire une argumentation, ce qui implique de choisir, en sus de certaines connexions rhétoriques, des arguments – i.e. du contenu – convaincants. L'objectif principal de cette communication est d'aborder ce problème de la double efficacité de la structuration de document à la lumière d'une application particulière : *la génération automatique de textes de preuves mathématiques*. La génération de preuves en langue naturelle fournit une illustration claire de l'acuité du problème en question. En effet, un objet preuve possède une structure formelle assez stable et bien définie, que l'on appelle l'arbre de preuve. D'un point de vue sémantique, une telle structure peut souvent se ramener à une série d'inférences (éventuellement imbriquées). On peut alors s'accorder l'intuition que l'information fournie par un arbre de preuve non seulement reflète le contenu sémantique du texte-preuve, mais donne aussi certaines indications sur sa structure rhétorique. C'est d'ailleurs ce sur quoi s'appuient un certain nombre de travaux en génération automatique de preuves (e.g. (Huang & Fiedler, 1997; Fiedler, 2001)) qui élaborent des stratégies plus ou moins sophistiquées d'exploitation ou de parcours de l'arbre pour produire, par appariements et regroupements, un plan de texte. De telles stratégies permettent de générer des textes corrects et conformes à la preuve formelle; cependant beaucoup sont stéréotypés, peu naturels, répétitifs, fastidieux, et parfois même abscons. Nous pensons que ces imperfections peuvent s'expliquer par le fait qu'à partir d'une preuve formelle, il n'est pas forcément trivial de distinguer ce qui devrait relever de la détermination de contenu vs. de la structuration de document. De plus même si la structure de l'arbre de preuve contraint fortement la structure rhétorique du texte, elle ne la détermine pas complètement. Empiriquement on constate que, souvent, l'organisation apparente d'un texte-preuve réel rend compte de la structure de l'arbre mais aussi des mécanismes de raisonnement utilisés dans la démonstration.

Dans cette communication nous allons proposer une stratégie originale de planification de textes de preuve, visant à obtenir des discours plus naturels et plus proches des preuves rédigées en langue naturelle. Les grandes lignes de cette stratégie seront exposées à travers la présentation du système GEPHOX qui est un générateur de textes de preuves mathématiques obtenues avec le système d'aide à la preuve PHOX (Raffalli & Roziere, 2002). Nous nous interessons ici au module *Quoi-dire?* du générateur dont l'organisation respecte une architecture standard en deux sous-modules : `ContDet` qui calcule le contenu à exprimer à partir de sorties de PHOX et en fonction de connaissances de l'utilisateur, et `DocStruct` qui calcule les plans du discours. Les plans de discours produits sont représentés dans le formalisme de la SDRT (Asher & Lascarides, 2003). Ce choix est motivé par deux raisons : d'abord nous nous fondons en grande partie sur le modèle de structuration de document de (Danlos *et al.*, 2001) qui montre l'efficacité de la SDRT pour la génération profonde ; ensuite nous adoptons les conclusions de (Zinn, 1999) selon lesquelles la DRT (et implicitement la SDRT) propose un formalisme particulièrement approprié pour l'analyse et la représentation des textes mathématiques.

2 Détermination du contenu : le module `ContDet`

Le système d'aide à la preuve PHOX permet de réaliser sur ordinateur des preuves mathématiques, en en garantissant la validité. Même si le logiciel dispose d'un algorithme de preuve automatique, son utilisation principale – e.g. pour l'enseignement des mathématiques – est la vérification des étapes de raisonnement. L'utilisateur *guide* PHOX pour démontrer des théorèmes mathématiques, en laissant le soin au logiciel de réaliser les vérifications et les opérations fastidieuses de la preuve.

GEPHOX prend en entrée des informations de deux types: le *script de preuve* qui représente la *trace* de la démonstration (i.e. les commandes entrées par l'utilisateur) et la *sortie de PhoX*

proprement dite, qui est constituée de fragments de l'*arbre de preuve* et que l'on peut donc voir comme le *contenu de la preuve*. Le module `ContDet` doit calculer à partir de cette entrée, et en tenant compte des connaissances de l'utilisateur, le message qui sera exprimé par le générateur. Dans GEPHOX les informations sont décrites dans les *concepts* et les *rôles* d'une logique de description (DL, cf. (Baader *et al.*, 2003)). Les connaissances sur le domaine et les connaissances propres de l'utilisateur sont respectivement représentées dans deux bases : DKB et UKB (UKB est en fait un sous-ensemble de DKB). Les bases de connaissances sont structurées en deux parties. D'une part, la **T-Box** (pour connaissances *t*erminologiques) encode les connaissances intensionnelles, *i.e.* les *concepts* et les *relations* du domaine. Par exemple, le concept `Entier` désigne l'ensemble des entiers naturels. D'autre part, la **A-Box** (pour connaissances *a*ssertionnelles) encode les connaissances extensionnelles, *i.e.* les individus de notre univers. Par exemple $n \in$ `Entier` introduit un individu appartenant au concept `Entier`.

Le principe général du calcul de contenu est le suivant. Il s'agit, dans un premier temps, de construire les expressions conceptuelles (**T-Box**) correspondant à l'entrée, en se rappelant des individus $\mathcal{I}$ qui leurs correspondent. C'est sur cette **T-Box** qu'opère la détermination de contenu proprement dite. Enfin, les expressions conceptuelles sont instanciées en utilisant les individus $\mathcal{I}$ pour produire la **A-Box** qui représente, sous formes de faits, le message à générer. Le module `ContDet` commence donc par *traduire* l'entrée du générateur en DL, en utilisant les concepts et rôles de la DKB. Si des définitions et des théorèmes issus de PHOX n'existent pas dans la base de connaissance, la procédure de traduction crée de nouveaux concepts en fonction de ceux déjà présents. Une fois construit l'ensemble des expressions conceptuelles ($\mathcal{EC}_D$) représentant l'entrée, `ContDet` doit *sélectionner* dans cet ensemble *ce qui doit être dit*. Sont alors mis en jeu des opérations de filtrage et de regroupement. Une première consiste à détecter des stratégies de raisonnement et leurs paramètres: par exemple, reconnaître l'annonce d'un raisonnement par cas; et ensuite calculer les différents cas et les associer à cette annonce. La deuxième opération est la mise en évidence des similitudes entre portions de preuve. On utilise à cet effet l'unification de concepts (Baader & Küsters, 2001). Enfin, `ContDet` cherche à apparier des définitions de concepts complexes de la DKB avec des expressions de $\mathcal{EC}_D$. Cela permet de synthétiser des groupes d'informations plus ou moins simples sous le chef d'un concept prédéfini (une telle opération est généralement dénommée *agrégation* en génération automatique). Les axiomes (i.e. les définitions de DKB) utilisés à cet effet sont gardés en mémoire pour établir ensuite des relations de second ordre sur les éléments de $\mathcal{EC}_D$. La **T-Box** ainsi obtenue donne une représentation de la preuve qui fait usage de tous les concepts du domaine[2]. La dernière étape de la détermination du contenu est l'instanciation de $\mathcal{EC}_U$ par les individus de $\mathcal{I}$ et la vérification de consistance du fragment de **A-Box** obtenu.

3 Structuration de document

La figure 1 donne un exemple de sortie de `ContDet`. Ce type de structure peut être vu comme un ensemble *ordonné*[3] de formes logiques qui marquent les pas significatifs de la démonstration. En regard de certaines formules figurent les axiomes qui ont permis de construire la formule.

[2]Pour produire un message coopératif et personnalisé, il est nécessaire de confronter cette représentation avec les connaissances de l'utilisateur (UKB), afin de vérifier s'il comprend tout ce que comporte $\mathcal{EC}_D$. Cela revient en fait à *expliquer* (i.e. décomposer) les concepts de $\mathcal{EC}_D$ qui ne figurent pas dans UKB. On obtient ainsi un nouvel ensemble d'expressions conceptuelles $\mathcal{EC}_U$, qui est calculé par *projection*de $\mathcal{EC}_D$ dans UKB.

[3]L'ordre provient de la sémantique du conjoncteur dynamique $\wedge$ qui n'est pas symétrique. Dans la figure 1, l'énumération des formules n'est donnée (pour l'instant) que pour la lisibilité du tableau.

De plus, la structure est plus riche qu'un simple arbre de preuve, puisque `ContDet` fait en sorte qu'elle contienne également des élements d'information provenant du script de preuve. Un des points centraux de notre stratégie de génération profonde va consister à exploiter cette richesse, notamment en postulant que les étapes du script traduites en formes logiques s'assimilent à des *intentions communicatives*. Ces intentions vont permettre en particulier de générer des actes de langage qui « humanisent » le texte de preuve en ajoutant des éléments de rhétorique autres que les habituelles relations logiques.

	Formes logiques	*Axiomes*
A	`sous-ensemble`(Q,$\mathbb{N}^*$) $\wedge$ **`define`**(Q,"$\exists n \in \mathbb{N}\ (m^2 = 2 * n^2)$")	`Q` $\doteq$ `Ensemble` $\wedge$ $\exists$`sous-ensemble`.{$\mathbb{N}^*$} $\wedge$ $\exists$`eq-def`.{$\exists n \in \mathbb{N}\ (m^2 = 2 * n^2)$}
B	`Entier`(m) $\wedge$ `neg-Q`(m)	`neg-Q` $\doteq \neg$ `Q`
C	$t_1 =$ `not-in`($sqrt-2$,$\mathbb{Q}$) $\wedge$ `Theoreme`(t_1) $\wedge$ **`annonce`**(t_1)	
D	1. `Entier`(m) $\wedge$ `Entier`(n) $\wedge$ **`choose`**(m) $\wedge$ **`choose`**(n) 2. $p_1 =$ `Prop`("$m^2 = 2 * n^2$") $\wedge$ **`suppose`**(p_1) 3. $l_1 =$ `lemme`("$lemme1$") 4. `neg-Q`(m) $\wedge$ **`implies`**(e_3,e_4) 5. `CaseReason`($current$) 6. `is-case`($current$,e_6) $\wedge$ `Nul`(m) 7. `Nul`(n) $\wedge$ **`implies`**(e_6,e_7) 8. `is-case`($current$,e_8) $\wedge$ `Entier-non-nul`(m) ...	...

Figure 1: Fragment du message pour «$\sqrt{2}$ *n'est pas rationnel* » (sortie de `ContDet`)

Le calcul du plan du texte est pris en charge par le module `DocStruct`, qui s'inspire de (Danlos *et al.*, 2001). Partant, `DocStruct` a deux tâches principales à mener pour produire une structure de document : 1) choisir les unités minimales de la structure, 2) choisir les relations rhétoriques qui lient les unités entre elles pour garantir la cohésion et la cohérence du discours.

La structure de la **A-Box** calculée par `ContDet` est celle d'un graphe connexe[4] (les sommets correspondant aux individus et les arcs à des rôles ou relations conceptuels). Dans ce graphe, certains sous-graphes sont fortement connexes (pseudo-cliques). Notre stratégie est de considérer que ce sont ces sous-graphes qui donneront lieu à des unités de discours, i.e. des segments minimaux de contenu. Par ailleurs, nos sous-graphes sont reliés entre eux par des dépendances que nous assimilerons à des relations d'ordre supérieur portant sur les unités de discours. Une telle relation correspond en fait soit à une commande du script, soit au résultat d'un axiome (e.g. *déduction, conclusion...*). Un résultat de segmentation apparaît dans la disposition de la figure 1 (colonne de gauche), où chaque ligne correspond à un segment de discours. Comme annoncé *supra*, notre futur plan de discours sera formalisé dans le cadre de la SDRT, i.e. sous forme d'une SDRS (Asher & Lascarides, 2003). Une SDRS est une structure dans laquelle des constituants de discours sont connectés par des relations rhétoriques. Les constituants sont des représentations sémantiques dynamiques héritées de la DRT, à savoir des DRS. Les segments de la figure 1 sont donc destinés à être traduits en DRS, et en accord avec (Danlos *et al.*, 2001), cette opération se fait conjointement à la sélection des relations rhétoriques.

Le modèle proposé par (Danlos *et al.*, 2001) est *déclaratif*. L'idée est que les relations rhétoriques sont associées à des postulats de sens et que ceux-ci sont considérés comme des conditions (des déclencheurs) de sélection d'une relation valide. Les conditions sont formulées dans le même langage que celui des formes logiques d'entrée pour permettre des appariements

[4]Nous n'allons nous intéresser ici qu'au cas d'un seul graphe connexe; si nous avons un graphe composé de plusieurs parties non connexes, le module `DocStruct` les traitera séparement. Dans notre message (Fig. 1) les parties A, B, C et D sont les représentants de sous graphes non connexes de notre entrée.

directs. La figure 2(a) illustre un tel appariement avec une règle pour la relation *Resultat*, qui en SDRT permet d'exprimer la causalité entre deux phrases. La règle a la forme suivante: *conditions* $\rightarrow$ SDRS. La SDRS de la partie droite est une portion de discours dans laquelle est instanciée la relation rhétorique déclenchée[5]. La SDRS obtenue en Fig. 2(a) pourra être générée en « Phrase$_1$. *Donc* Phrase$_2$. »[6].

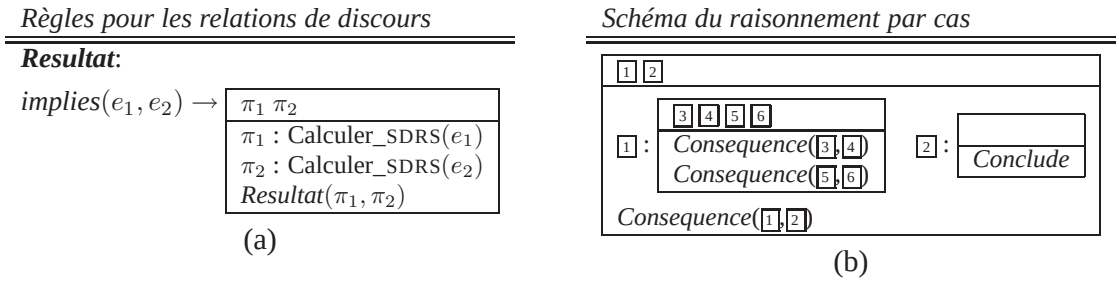

Figure 2: Règles et Schémas de Structuration Rhétorique

Cependant notre stratégie se distingue de (Danlos *et al.*, 2001) sur plusieurs points. D'abord les formes logiques de notre entrée comportent des informations qui peuvent refléter des intentions communicatives. Cela permet de sélectionner dans un plus large éventail de relations rhétoriques, notamment en faisant varier les forces illocutoires. Nous reprenons en effet l'hypothèse de (Asher & Lascarides, 2003) selon laquelle les arguments des relations rhétoriques sont des actes de langage et que les relations induisent des typages illocutoires. Il s'agit ici pour nous de traduire des « actes de démonstration » (les commandes saisies dans PHOX) en actes de langage. Les règles sont similaires à celles de Fig. 2(a) mais les relations déclenchées pourront spécifier une force illocutoire particulière. Par exemple, une condition comme *choose*(n) pourra ainsi donner lieu à une assertion (« on choisit n... ») ou un impératif (« soit n... »). Nous postulons aussi qu'un impératif déclenche une relation qui a portée sur tout le reste du discours[7]; en l'occurence, pour les textes de preuve, nous posons la relation *Background*$_i$ pour « impératif d'arrière-plan ».

Une seconde particularité de notre approche tient au traitement des annonces de stratégies de raisonnement mentionnées dans la forme logique et issues du script de preuve. Une telle annonce ne déclenche pas directement une relation rhétorique, mais un *schéma* rhétorique propre au raisonnement en question. Un schéma est en fait une structure contenant plusieurs relations. Par exemple, le schéma du raisonnement par cas est illustré en Fig. 2(b). Ce type de traitement s'explique par le fait que ces stratégies de raisonnement s'appuient sur des théorèmes fondamentaux (e.g. le tiers exclu pour le raisonnement par cas) qui en soi ne sont jamais explicités dans la démonstration, mais qui en revanche impliquent des agencements discursifs bien précis (e.g. une suite de « *si...*, *alors...* » dans le raisonnement par cas). Un schéma ainsi enclenché ne contient qu'un squelette rhétorique, qui sera paramétré ensuite en fonction du contenu de la forme logique.

[5]La fonction `Calculer_SDRS` fait partie de la procédure de structuration et permet de construire le discours récursivement.

[6]Notons que comme dans (Danlos *et al.*, 2001), une condition comme *implies*(e_1, e_2) peut donner lieu à autre chose qu'une relation rhétorique, par exemple un prédicat (« verbal ») qui fonde une DRS (« X_1 implique X_2. », « de X_1 on obtient X_2. » ou « X_2 se déduit de X_1 »).

[7]Ce que nous ne démontrerons pas ici pour des raisons de place.

4 Conclusion

Le modèle de génération profonde en cours d'implémentation de GEPHOX[8] propose une stratégie de planification originale sur au moins deux aspects. D'abord, par rapport aux générateurs automatiques de textes de preuves déjà existants, les (plans de) discours ici produits s'annoncent plus naturels et plus proches des preuves rédigées manuellement. Cela est dû à la spécificité du module `ContDet` qui exploite avantageusement les sorties du prouveur pour *engendrer dynamiquement*[9] une structure linguistique spécifique et distincte, notamment en dégageant des informations de nature intentionnelle (e.g. les commandes du script). Par ailleurs, si l'utilisation d'intentions n'est pas neuve en génération (e.g. cf. (Moore & Paris, 1993)), la mise en œuvre que nous présentons se distingue par la simplicité du traitement. En effet l'approche est hybride en ce sens que les conditions intentionnelles et les conditions informationnelles (sémantiques) de l'entrée donnent toutes des structures de discours de même type: des (portions de) SDRS. Celles-ci peuvent ensuite s'assembler selon un procédé unique (qui dépend seulement des contraintes de bonne formation stipulées en SDRT). Cela permet, entre autres, de s'affranchir de la complexité, souvent mentionnée, d'une gestion séparée des buts communicatifs et de leurs interactions parfois discordantes avec les structures rhétoriques.

Références

ASHER N. & LASCARIDES A. (2003). *Logics of Conversation*. Cambridge: CUP. (à paraître).

BAADER F. & KÜSTERS R. (2001). Unification in a Description Logic with Transitive Closure of Roles. In R. NIEUWENHUIS & A. VORONKOV, Eds., *Proceedings of LPAR 2001*, volume 2250 of *Lecture Notes in Artificial Intelligence*, Vienna: Springer–Verlag.

F. BAADER, D. L. MCGUINNESS, D. NARDI & P. F. PATEL-SCHNEIDER, Eds. (2003). *Description Logics Handbook: Theory, Implementation and Applications*. Cambridge University Press.

DANLOS L., GAIFFE B. & ROUSSARIE L. (2001). Document structuring à la SDRT. In *Proceedings of the 8th European Workshop on Natural Language Generation (EWNLG'2001)*, p. 11–20, Toulouse.

FIEDLER A. (2001). *User-adaptive proof explanation*. PhD thesis, Naturwissenschaftlich-Technische Fakultät I, Universität des Saarlandes, Saarbrücken, Germany.

HALLGREN T. & RANTA A. (2000). An extensible proof text editor. In M. PARIGOT & A. VORONKOV, Eds., *Proceedings of LPAR'2000, LNCS/LNAI 1955*, p. 70–84, Heidelberg: Springer Verlag.

HUANG X. & FIEDLER A. (1997). Proof verbalization as an application of NLG. In *IJCAI'97 Proceeding (2)*, p. 965–972.

MOORE J. D. & PARIS C. L. (1993). Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, **19**(4), 651–694.

RAFFALLI C. & ROZIERE P. (2002). *The PhoX Proof checker documentation*. LAMA, Université de Savoie / Université Paris 7.

REITER E. & DALE R. (2000). *Building Natural Language Generation Systems*. Studies in Natural Language Processing. CUP.

ZINN C. (1999). Understanding mathematical discourse. In *Proceedings of Amstelogue'99, 3rd Workshop on the Semantics and Pragmatics of Dialogue*, Amsterdam.

[8]Pour des raisons de place, nous n'avons pas détaillé ici l'algorithme complet de planification de GEPHOX, préférant concentrer notre propos sur les principes qui sous-tendent la structuration de document.

[9]Et en cela, nous nous distinguons de l'approche par traduction (*mapping*) de (Hallgren & Ranta, 2000).