EASYTEXT : un système opérationnel de génération de textes

Frédéric Meunier[1] Laurence Danlos[2] Vanessa Combet[1]
(1) Watch System Assistance
(2) Université Paris Diderot, ALPAGE
frederic.meunier@watchsystance.com, laurence.danlos@linguist.jussieu.fr,
vanessa.combet@watchsystance.com

EASYTEXT est un système de génération de textes opérationnel auprès de Kantar Media, une filiale de TNS-Sofres. Cette société compile pour ses clients des données chiffrées sur leurs investissements publicitaires et envoie à chaque client sept tableaux tous les mois, comme celui de la Figure 1. Avant EASYTEXT, ces tableaux étaient accompagnés d'un commentaire général rédigé par un chargé d'étude. Le besoin s'est fait sentir d'assortir ce commentaire général de commentaires spécifiques à chaque tableau. La charge de rédaction étant alors trop lourde pour les chargés d'étude, l'idée a surgi de faire générer ces commentaires spécifiques par un système automatique.

EASYTEXT repose sur le formalisme G-TAG, un formalisme lexicalisé reposant sur les grammaires d'arbres adjoins (TAG) (Danlos, 1998). Ce formalisme a été étendu en amont pour les tâches de détermination du contenu et de structuration du texte, en suivant l'architecture décrite dans (Danlos & Ghali, 2002). La détermination du contenu revient à surligner certaines cellules du tableau. Cette tâche a été guidée par les règles métier indiquées par les chargés d'étude de TNS-Sofres. La structuration du texte consiste à introduire des relations rhétoriques (e.g. $Contraste$ ou $Parallèle$) entre le contenu sémantique des cellules surlignées, voir (Danlos *et al.*, 2001)).

G-TAG avait été implémenté dans les années 90' en ADA par F. Meunier (1997). Celui-ci a re-implémenté G-TAG (avec les extensions en amont décrites ci-dessus) en .NET, ce qui permet à EASYTEXT d'être intégré au système d'information de TNS-Sofres. La génération d'un commentaire comme celui de la Figure 1 demande en moyenne 400ms de CPU-Times. Cette implémentation intègre des outils ergonomiques pour renseigner les bases lexicales TAG et pour visualiser les différentes structures arborescentes dynamiquement construites ou en cours de construction.

Les bases lexicales ont été renseignées par une linguiste, V. Combet, qui a travaillé en étroite collaboration avec les chargés d'étude de TNS-Sofres. Une attention particulière a été portée sur la variation linguistique afin de ne pas produire des textes monotones qui auraient lassé les clients de TNS-Sofres. Cette variation concerne principalement les choix lexicaux (e.g. *augmenter, être en (forte/moyenne/faible) augmentation/hausse, (presque/plus que) doubler/tripler*) et l'ordre des syntagmes à position plus ou moins libre.

TNS-Sofres a été satisfait des résultats de EASYTEXT (même au delà de ses espérances), et donc commercialise ce service à ses clients depuis Avril 2010. EASYTEXT est donc un des tous premiers systèmes de génération de textes opérationnel en France, et il en existe peu en dehors de l'hexagone.

Figure 1 : Exemple de tableau et de commentaire généré automatiquement

Références

DANLOS L. (1998). G-TAG : un formalisme lexicalisé pour la génération de textes inspiré de TAG. *Revue TAL*, **39**(2).

DANLOS L., GAIFFE B. & ROUSSARIE L. (2001). Document structuring à la SDRT. In *International workshop on text generation - ACL*, p. 94–102, Toulouse.

DANLOS L. & GHALI A. E. (2002). A completed and integrated NLG system using NLU and AI tools. In *Proceedings of COLING'02*, Tapei, Taiwan.

MEUNIER F. (1997). *Implémentation du formalisme G-TAG*. Thèse de doctorat en informatique, Université Denis Diderot, Paris 7.