Un segmenteur de texte en phrases guidé par l'utilisateur

Thomas Heitz*

*Université Paris-Sud XI, 91405 Orsay CEDEX
heitz@lri.fr,
http://www.lri.fr/∼heitz

Résumé. Ce programme effectue une segmentation en phrases d'un texte. Contrairement aux procédures classiques, nous n'utilisons pas d'annotations préliminaires et tirons parti d'un apprentissage guidé par l'utilisateur.

La segmentation en phrases entièrement automatisée et avec une importante proportion des corpus annotés en phrases manuellement est déjà très efficace. De même, la segmentation en phrases à l'aide de dictionnaires et de règles syntaxiques spécifiquement adaptées à un corpus donné est aussi relativement efficace.

Ce qui nous intéresse ici est donc la segmentation d'un corpus en phrases sans aucune segmentation initiale et avec l'aide de l'utilisateur pour diriger les traitements et notamment l'apprentissage. Ce que nous appelons apprentissage guidé. Le but est de minimiser le temps consacré par l'utilisateur à annoter des fins de phrases. C'est pourquoi nous utilisons au maximum les connaissances générales de l'écriture du langage naturel et nous présentons à l'utilisateur les seuls cas les plus ambigus.

Le but est d'annoter le mot précédent et suivant de chaque point suivi d'un espace afin de déterminer si la phrase doit être terminée sur ce point ou non.

L'idée qui est utilisée dans ce segmenteur est la suivante. Le mot précédent le point peut être une abréviation et dans ce cas il est fort probable que le point ne soit pas une fin de phrase. Le mot suivant le point peut être un mot toujours capitalisé, c'est-à-dire commençant par une majuscule dans tout le texte, et dans ce cas il est fort probable que le point ne soit pas une fin de phrase.

Les **annotations utilisées** pour classer les mots précédents et suivants les points suivis d'un espace sont les annotations *certain* et *impossible* qui correspondent aux mots que l'utilisateur considère comme étant (respectivement n'étant pas) certainement une abréviation terminée par un point ou un mot toujours capitalisé. L'annotation *possible* correspond aux éléments indéterminés qui deviendront *certain* ou *impossible* ultérieurement.

La procédure globale de segmentation se déroule selon les étapes suivantes :

① Établissement de **statistiques** sur les abréviations probables et les mots capitalisés probables sur le corpus complet. Notamment le nombre d'occurrences avec et sans point final et avec et sans majuscule initiale.

② **Annotation automatique** sur un extrait du corpus des abréviations à l'aide de listes de mots communs, d'abréviations et de règles syntaxiques. L'utilisateur peut choisir d'avoir des résultats plus précis sur les annotations *certain* et *impossible* mais obtiendra en contrepartie une plus grande quantité d'annotations *possible*. L'utilisateur peut ensuite classer les abréviations restées *possible* en *certain* et *impossible*.