Un outil de géolocalisation et de résumé automatique pour faciliter l'accès à l'information dans des corpus d'actualité

Emilie Guimier De Neef, Aurélien Bossard, Frédéric Gavignet, Olivier Collin

Orange Labs R&D
2 av Pierre Marzin, 22307 Lannion CEDEX
prenom.nom@orange-ftgroup.com,

1 Introduction

Face à l'abondance de contenus d'actualité et à leur continuel renouvellement, le défi pour les services d'agrégation de news est de parvenir à valoriser ces contenus auprès des utilisateurs sans les noyer d'informations au moyen de techniques rapides et automatiques.

Les contenus de presse sont généralement classés dans des catégories (sport, culture, économie...), ce qui permet un accès thématique à l'actualité. L'annotation des contenus par des méthodes de TALN ouvre la porte à de nouvelles modalités d'accés à l'actualité comme l'accès géolocalisé à l'actualité, ce que se propose d'illustrer notre premier démonstrateur.

Les techniques de clustering regroupent les articles qui parlent des mêmes actualités et offrent à l'utilisateur un accès par sujet (voir par exemple le service www.2424actu.fr). On se trouve alors face à des contenus redondants dont il s'agit de synthétiser l'information pour l'utilisateur. Nous proposons un module de résumé automatique multi-documents qui réalise une extraction des phrases les plus importantes en maximisant l'information et minimisant la redondance informationnelle (Bossard, 2009).

2 Géolocalisation

Les dépêches de presse et articles issus de la plate-forme 2424actu sont des contenus courts et thématiquement homogènes regroupés en sujets par une technique de clustering entièrement automatique. La brique de géolocalisation présentée ci-dessous permet un accès géolocalisé aux clusters (cf figure...). Une fonctionnalité de zoom permet d'affiner la granularité et de filtrer les lieux en fonctions d'un continent, d'une région etc.

La géolocalisation des clusters se fait en trois étapes. Tout d'abord, chaque contenu fait l'objet d'une extraction d'entités nommées (repérage des personnes, lieux, organisations) au moyen d'une technologie symbolique à base de dictionnaires et de règles syntaxiques (Heinecke et al., 2008) (*cf* Fig. 2).

Pour chaque news, les indicateurs linguistiques locatifs (pays, continents, départements, régions, villes, micro-toponymes...) sont extraits.