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.

FIG. 1 – *Interface de géolocalisation de 2424Actu*

Le préfet du<LOC>Var</LOC> vient d'être informé par l<ORG>Agence Régionale de Santé</ORG> (<ORG>ARS</ORG>) de<LOC>Provence Alpes Côte d Azur</LOC> que le premier cas autochtone de chikungunya vient d être diagnostiqué dans le<LOC>Var</LOC>, à<LOC>Fréjus</LOC>, sur une enfant de 12 ans, actuellement suivie à son domicile ", a annoncé la préfecture dans ce communiqué, précisant qu' il s' agissait " pour l instant d un cas isolé ". (...) Le préfet du<LOC>Var</LOC>,<NOMPERS>Hugues Parant</NOMPERS>, fera un point de la situation <TIME>lundi à 14 h 30</TIME> en préfecture, lors d' une conférence de presse. Depuis le début de l' été, le moustique - tigre " aedes albopictus ", vecteur potentiel du chikungunya et de la dengue, fait l'objet d' une surveillance particulière dans trois départements de<LOC>Provence - Alpes - Côte d'Azur</LOC> [...]

FIG. 2 – *Exemple d'un texte automatiquement annoté en entités nommées.*

Ensuite, les coordonnées géographiques de ces différents lieux sont récupérés par interrogation de la base Geonames (www.geonames.org). Au final, les informations locatives sont compilées et hiérarchisées en une forêt pondérée d'hypothèses.

Finalement, les hypothèses locatives des textes d'un cluster sont agrégées et confrontées pour décider d'une localisation unique pour le cluster, si possible sous forme d'un triplet continent/pays/ville.

3 Résumé automatique multi-documents

Les clusters de 2424actu regroupent les dépêches en sujets. L'objectif est de proposer à l'utilisateur une synthèse de l'information du cluster exhaustive, pertinente et non redondante.

Nous proposons une approche fondée sur le regroupement des phrases en classes sémantiques (Bossard, 2009), qui constituent des sous-thèmes. Les phrases sont regroupées selon leurs similarités grâce à un algorithme de clustering, *fast global k-means*. L'idée est alors d'extraire une phrase par sous-thème afin d'obtenir le résumé le plus diversifié possible. Afin de déterminer la phrase à extraire, nous utilisons une combinaison de deux scores de centralité : la centralité globale, utilisée dans tous les systèmes de résumé automatique existants, et la centralité locale, propre aux sous-thèmes. Il s'agit donc d'évaluer quelle est la phrase la plus centrale du point de vue du contenu global des documents, mais également dans sa classe.

Documents : Les routiers se mobilisent aussi - La CFDT appelle les routiers à la grève - La CFDT appelle les routiers à l'action - Les routiers prêts à rejoindre la contestation - Les routiers n'excluent pas de stopper la distribution de denrées alimentaires - Retraites : La CFDT appelle les routiers à la grève - Les routiers appelés à se joindre au mouvement de blocages - Les routiers se mobilisent : "Une action de solidarité avec le reste de la population" - Les chauffeurs routiers annoncent des actions - Et maintenant, les routiers ? - RÉFORME DES RETRAITES - Les routiers annoncent des blocages et opérations "escargot" - [...]
Résumé : La réforme des retraites entre dans une semaine décisive avec une nouvelle journée d'action mardi, à la veille du vote du projet de loi sous tension au Sénat, tandis que plane toujours la menace d'une pénurie de carburant et d'un durcissement du mouvement chez les routiers. Le premier syndicat du transport routier, la CFDT, appelle les salariés à organiser des barrages filtrants ou des opérations escargots. Ils ont entamé des opérations escargots sur plusieurs axes routiers. Des véhicules personnels ou loués par les syndicats car les routiers ne peuvent pas utiliser leur camion professionnel.

FIG. 3 – *Exemple d'un résumé : titres des documents et résumé en 100 mots maximum.*

Les expériences menées sur un corpus d'évaluation français développé dans le cadre du projet RPM2[1] montrent que notre approche surpasse les techniques classiques, telles que MMR, Centroïde et LexRank (voir Figure 3 un aperçu d'une classe événementielle et son résumé).

4 Conclusion

Nous présentons des démonstrateurs qui illustrent deux fonctionnalités qui facilitent l'accès à l'information. La géolocalisation offre un accès cartographique à l'actualité mais d'autres modalités sont également envisageables (accès par les personnalités...). Dans tous les cas, l'enjeu est la qualité de l'annotation des contenus et en particulier celui de la désambiguïsation : distinguer Matignon, commune des Côtes d'Armor, de Matignon, résidence du premier ministre etc.

En matière de filtrage et/ou résumé de l'information, la problématique est celle de la représentation de l'information. La plupart des approches projettent les textes sur les mots ou les lemmes qu'ils contiennent, ce qui reflète forcément imparfaitement la sémantique du texte. Une utilisation plus fine des informations rhétoriques du texte et sémantiques des énoncés reste une piste à explorer pour améliorer les performances des systèmes de résumé.

Références

Bossard, A. (2009). CBSEAS, a new approach to automatic summarization. In *SIGIR 2009 Conference - Doctoral Consortium*, Boston, USA.

Heinecke, J., G. Smits, C. Chardenon, E. Guimier De Neef, E. Maillebuau, et M. Boualem (2008). Tilt : plate-forme pour le traitement automatique des langues naturelles. *TAL 2*(49).

Summary

In this paper, we present two tools developped by Orange Labs in order to facilitate content access.The first one offers a map interface where news are geolocalized. The second one provides an overview of a news topic through a summarization tool.

[1] http://labs.sinequa.com/rpm2/projet.html