Faciliter les contributions personnelles pour préserver la mémoire des événements historiques

Pierrick Bruneau, Olivier Parisot, Thomas Tamisier

LIST, 5 Avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette
pierrick.bruneau@list.lu, http://www.list.lu

Résumé. Un aspect essentiel dans la préservation du patrimoine culturel réside dans la collecte et l'assemblage des témoignages provenant de citoyens ordinaires. Dans cet article, nous présentons une architecture logicielle facilitant la saisie et le partage de témoignages concernant la période de la construction européenne au Luxembourg. En rédigeant son témoignage, l'utilisateur obtient les résultats d'une extraction de connaissances sur le contenu saisi, indiquant notamment des entités et informations liées.

1 Introduction

La collecte et l'assemblage de témoignages et anecdotes de citoyens ordinaires est un aspect important, mais sous-estimé, de la préservation du patrimoine culturel. Tandis que l'adoption massive des médias sociaux facilitera ce travail pour les générations futures, de telles données sont par exemple difficilement disponibles pour la période de construction européenne (environ 1945-1975).

Dans cet article, nous nous concentrons sur les traces de cette époque au Luxembourg. Des ateliers préliminaires ont été organisés avec des utilisateurs cibles (entre 75 et 85 ans). L'approche de la *sonde technologique* a été suivie (Hutchinson et al., 2003), en confrontant les utilisateurs à une variété de sources (e.g. images, livres, cartes). Les interactions entre utilisateurs et les sources ont été codifiées. Parmi nos observations, il apparait clairement qu'un effort initial est nécessaire pour stimuler le récit des personnes âgées. Celles-ci ont manifesté un intérêt certain pour les expériences multimédia et interactives, sous réserve de recevoir l'aide technique nécessaire. Le présent article vise à définir une architecture de données et de logiciels facilitant ce processus d'incitation. Nous envisageons la saisie semi-automatique d'histoires, où l'utilisateur édite librement son contenu, et reçoit une aide à partir d'une extraction de structure automatique.

Après un passage en revue de la littérature pertinente, nous décrivons une chaîne de traitement pour extraire la connaissance contenue dans un texte brut, utiliser le résultat d'extraction pour l'enrichir à partir de sources de données externes, et transformer cet agrégat en une forme propre à l'affichage. Notre approche implique la définition de structures de données adaptées au contenu narratif autobiographique. Des extraits de résultats obtenus à partir d'un petit corpus en Français fourni par le *Centre National de l'Audiovisuel* (CNA [1]) luxembourgeois sont présentés. De nombreuses perspectives, résumées en Section 4, sont ouvertes par ce travail.

1. http://www.cna.public.lu

2 Travaux connexes

L'analyse narrative a été appliquée à de nombreux domaines, e.g. en gestion de crise (Scherp et al., 2009), en connaissance de la situation (Van Hage et al., 2012), ou en gestion de documents historiques (Segers et al., 2011). (Scherp et al., 2009; Van Hage et al., 2012) définissent une taxonomie de liens entre événements (composition, causalité, corrélation, et documentation) pertinente dans notre contexte. Toutefois le niveau d'abstraction utilisé n'autorise pas un vocabulaire contrôlé de prédicats. Pareillement à (Van der Meij et al., 2010; Segers et al., 2011), les auteurs mettent l'accent sur l'interopérabilité entre ontologies. Des rôles (ou *facettes* dans (Mulholland et al., 2012), e.g. acteur, date, lieu) s'appliquant aux événements sont explicitement définis dans (Segers et al., 2011). Ce formalisme convient aux événements historiques au sens large (e.g. la Révolution Française dans (Segers et al., 2011)), mais pas à une narration autobiographique. Notons que l'association de bornes temporelles aux faits d'une ontologie générique a aussi été considérée, e.g. dans YAGO2 (Hoffart et al., 2013).

Les contributions de (Zarri, 2009) sont les plus clairement liées à notre travail. Un vocabulaire contrôlé de prédicats et de liens adapté à l'analyse de la narration non-fictionnelle y est défini. Plutôt que le terme d'*histoire*, sujet à confusion, ils définissent la *fabula* en tant qu'un ensemble d'événements et de faits. L'*intrigue* ajoute des liens logiques et chronologiques entre événements. La *présentation* concerne la forme dans laquelle les intrigues sont montrées. D'autres travaux en analyse narrative s'intéressent à l'association entre des histoires arbitraires et les structures narratives classiques (Tilley, 1992; Yeung et al., 2014). Dans notre contexte, nous pouvons avoir affaire à des anecdotes, *a priori* difficiles à associer à de telles structures. Des propriétés plus abstraites, comme les sentiments attachés à une histoire, ont aussi été extraites dans (Min et Park, 2016), puis utilisées pour analyser la structure de livres.

Le processus d'association automatique d'un texte arbitraire à une taxonomie de types d'entités et de prédicats est rarement considéré dans la littérature. Certains travaux supposent explicitement que ce processus doit être réalisé manuellement (Mulholland et al., 2012), ou de manière participative (Bollacker et al., 2008). La structure des pages Wikipedia a été exploitée par (Suchanek et al., 2008). De manière alternative, une heuristique à base de termes-clés est utilisée par (Gaeta et al., 2014) afin de déterminer les liens entre événements. Les techniques de *Traitement Automatisé de la Langue naturelle* (TAL), telles que la *Reconnaissance d'Entités Nommées* (REN) ont été utilisées par (Segers et al., 2011; Van Hooland et al., 2015) pour extraire des faits et des événements. Notons que les entités dans les modèles d'événements tels que SEM (Van Hage et al., 2012) sont proches des types extraits par les méthodes de REN (e.g. personnes, lieux, dates (Favre et al., 2005)).

3 Architecture proposée et résultats expérimentaux

Nous décrivons une architecture logicielle facilitant l'extraction de faits, d'événements, et d'intrigues depuis du texte brut (Figure 1). Initialement, l'utilisateur peut s'inspirer en consultant les faits stockés dans une base locale, et simplement commencer à saisir son histoire. Les faits, les événements et les liens entre événements sont construits à partir de ce contenu initial. Dans cette section, nous détaillons ce processus, que nous enrichissons de l'accès à des sources de données externes. Les moyens employés pour le stockage local sont également discutés.

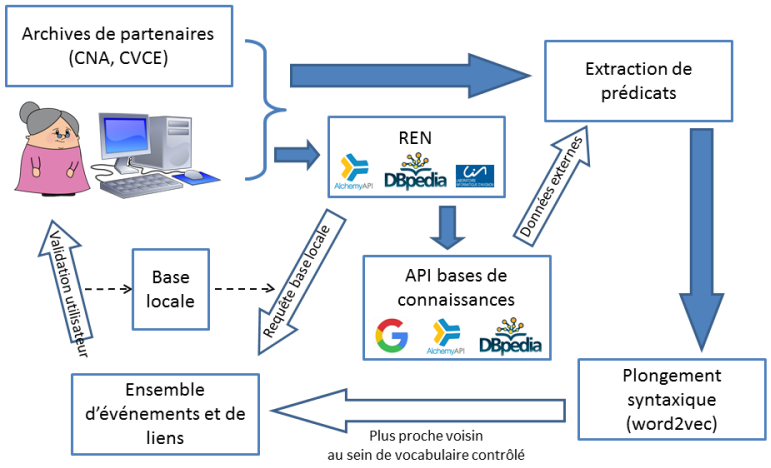

FIG. 1 – *Architecture logicielle facilitant la saisie et le rappel de témoignages.*

Suivant la terminologie définie par (Zarri, 2009) et évoquée en Section 2, la narration brute (i.e. le texte) est vue comme une forme de présentation. Le but de notre chaîne de traitement est d'extraire la *fabula* et l'intrigue sous-jacentes à cette représentation initiale, ainsi que suggérer des informations à lier à cette connaissance extraite. Des exemples de la terminologie de (Zarri, 2009) sont indiqués dans la table 1. Des *modèles* de prédicats à deux niveaux sont à la base de cette structure. Des entités sont associées à ces modèles par une nomenclature de *rôles*. Des combinaisons complexes d'entités et de liens entre événements peuvent être générées en utilisant des opérateurs d'*association*.

Modèles	**Rôles**	**Associations**
se comporter *(attitude)*, exister *(naissance)*, expérimenter *(interaction sociale positive)*, déplacer *(donner)*, posséder, produire *(refuser)*	sujet, objet, source *(responsabilité)*, bénéficiaire, mode *(instrument)*, thème	disjonctive, collective, énumérative, cause, référence, but, motif, condition

TAB. 1 – *Terminologie de (Zarri, 2009).*

Les entités nommées peuvent être extraites de texte en Français grâce à des API distantes telles qu'AlchemyAPI (AlchemyAPI, 2016), ou des outils comme LIA (Favre et al., 2005). Du texte et de l'information additionnels peuvent être obtenus en utilisant les noms extraits comme clés dans des bases de connaissance telles que le *Google Knowledge Graph* (GKG) (Google, 2012). Du contenu additionnel peut également être obtenu via les API de moteurs de recherche

classiques, en restreignant les résultats au domaine de Wikipedia, par exemple (Segers et al., 2011).

Des outils de TAL plus bas niveau, tels que le tagueur *Part-Of-Speech* (POS) disponible dans LIA, sont nécessaires pour extraire les prédicats associant les entités entre elles. Les prédicats extraits doivent ensuite être associés au vocabulaire contrôlé résumé dans la table 1. De manière alternative aux techniques mentionnées en section 2, nous proposons de guider l'association grâce à un plongement lexical, qui consiste à exploiter la structure syntaxique pour associer un vecteur numérique multidimensionnel à chaque mot de vocabulaire (Mikolov et al., 2013). Des candidats pertinents à l'association sont alors les plus proches voisins d'un mot arbitraire au sein du vocabulaire contrôlé. En d'autres termes, plutôt que d'utiliser une taxonomie exhaustive, nous exploitons la structure implicite au plongement lexical. Ce plongement est réalisable via des bibliothèques telles que TensorFlow (TensorFlow, 2016).

Par défaut l'association *référence* (i.e. causalité faible) peut être appliquée à la suite d'événements détectée. Les événements et les faits extraits à partir de sources externes peuvent être associés à l'entité qui a causé leur requête. Les références temporelles trouvées dans le texte peuvent aussi caractériser l'ensemble des événements extraits.

Dans une ontologie générique, (Hoffart et al., 2013) ont décomposé des faits complexes en faits simples grâce à la *réification*. Nous reprenons cette technique afin de dérouler les événements et les intrigues définis dans la section précédente. Ceci permet de se reposer sur les outils de stockage de triples RDF, tels que ceux implémentés dans la plateforme Drupal (Corlosquet et al., 2009; Drupal, 2011). Cette plateforme offre également un système avancé de profils utilisateurs, qui s'avéreront utiles au moment de connecter notre chaîne de traitement à des vues interactives.

Pour illustrer notre approche, nous avons adapté la description d'une vidéo fournie par le CNA : *Avec mon frère, nous étions en ville à l'occasion de la visite de Winston Churchill à Luxembourg les 14 et 15 juillet 1946. Il a été accueilli par le Prince Felix et le Prince Jean. Churchill s'est ensuite rendu à l'Hôtel de Ville où l'a reçu M. Hamilius.* En effectuant une REN (ici avec AlchemyAPI) sur ce texte on obtient les entités suivantes : *Winston Churchill, Prince Felix, Prince Jean, Hôtel de Ville, M. Hamilius*.

GKG a ensuite permis d'obtenir des informations additionnelles sur ces entités. Pour limiter les ambigüités, le mot-clé *Luxembourg* a été combiné aux requêtes. La description suivante a été obtenue pour *M. Hamilius* : *Émile Hamilius, né le 16 mai 1897 à Esch-sur-Alzette et mort le 7 mars 1971 à Luxembourg, est un footballeur et homme politique luxembourgeois.*

Notons que cette étape d'enrichissement a introduit de nouvelles entités (e.g. *Esch-sur-Alzette*). Les formes infinitives des prédicats détectés peuvent être extraites avec les outils LIA : *être, accueillir, être, rendre, recevoir*. Les prédicats et entités découverts peuvent alors être utilisés pour la construction d'événements.

4 Conclusion

Nous avons décrit une chaine de traitement qui combine des techniques de TAL et de modélisation de connaissances et d'événements afin d'améliorer le rappel de témoignages personnels. L'affichage interactif de son résultat, et la correction de ses erreurs par un utilisateur sont hors du champ d'étude de cet article. Outre l'intégration des composants décrits dans la section 3, ce sont les perspectives les plus immédiates de notre travail. Même si une variété

d'API et de librairies de TAL ont été testées sur des données réelles, dans un souci de concision seul un extrait est donné en section 3. Une évaluation approfondie des résultats selon les métriques usuelles (e.g. précision, rappel) devrait être incluse dans l'extension de ce travail.

Nous prévoyons également de traiter les conflits entre faits. L'utilisation classique de la déduction logique consiste à inférer de nouveaux faits (Suchanek et al., 2008), même si la détection de contradictions a déjà été abordée dans ce domaine (Paulheim, 2016). Les auteurs de Knowledge Vault (Dong et al., 2014) combinent l'apprentissage automatique, des heuristiques textuelles et des faits obtenus depuis Freebase (Bollacker et al., 2008) afin d'extraire des faits depuis des sources hétérogènes, et estimer leur exactitude. L'hypothèse du *Monde Fermé Local* qu'ils utilisent est une clé possible pour la détection de conflits entre faits : considérant un sujet s, un prédicat p et des objets o et o', si la base de faits contient à la fois les triplets (s, p, o) et (s, p, o'), un arbitrage pourrait être demandé à l'utilisateur.

Les *fabulae* et intrigues obtenues sont subjectives, i.e. elles contiennent des marqueurs syntaxiques réflexifs (e.g. moi, mon frère). Les systèmes de REN testés en section 3 ne sont pas capables de détecter ces marqueurs. Une heuristique simple pourrait être développée à partir d'une liste de mots-clés ou de tags POS. Des outils de résolution d'anaphore comme GUITAR pourraient aussi être testés (Poesio et Kabadjov, 2004).

Remerciements : Ce travail, réalisé pour le projet LOCALE et financé par le *Fonds National de la Recherche*, a utilisé des données du *Centre National de l'Audiovisuel* luxembourgeois. Nous voulons enfin exprimer nos vifs remerciements aux membres de l'association AMIPERAS ayant volontairement participé aux ateliers préparatoires.

Références

AlchemyAPI (2016). AlchemyAPI. `http://www.alchemyapi.com/`.

Bollacker, K. et al. (2008). Freebase : a collaboratively created graph database for structuring human knowledge. In *SIGMOD*, pp. 1247–1250.

Corlosquet, S., R. Delbru, T. Clark, A. Polleres, et S. Decker (2009). Produce and Consume Linked Data with Drupal ! In *International Semantic Web Conference*, pp. 763–778.

Dong, X. et al. (2014). Knowledge vault : A web-scale approach to probabilistic knowledge fusion. In *SIGKDD*, pp. 601–610.

Drupal (2011). Drupal modules. `https://www.drupal.org/project/`.

Favre, B., F. Béchet, et P. Nocéra (2005). Robust named entity extraction from large spoken archives. In *HLT/EMNLP 2005*, pp. 491–498.

Gaeta, A., M. Gaeta, et G. Guarino (2014). RST-based methodology to enrich the design of digital storytelling. In *IEEE INCOS 2015*, pp. 720–725.

Google (2012). Introducing the knowledge graph. `http://tinyurl.com/zofw8fb`.

Hoffart, J., F. M. Suchanek, K. Berberich, et G. Weikum (2013). YAGO2 : A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence 194*, 28–61.

Hutchinson, H. et al. (2003). Technology probes : inspiring design for and with families. In *SIGCHI*, pp. 17–24.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, et J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *NIPS*, pp. 3111–3119.

Min, S. et J. Park (2016). Mapping out narrative structures and dynamics using networks and textual information. *arXiv preprint arXiv :1604.03029*.

Mulholland, P., A. Wolff, et T. Collins (2012). Curate and storyspace : an ontology and web-based environment for describing curatorial narratives. In *ESWC 2012*, pp. 748–762.

Paulheim, H. (2016). Knowledge graph refinement : A survey of approaches and evaluation methods. *Semantic Web*, 1–20.

Poesio, M. et M. A. Kabadjov (2004). A general-purpose, off-the-shelf anaphora resolution module : Implementation and preliminary evaluation. In *LREC*.

Scherp, A., T. Franz, C. Saathoff, et S. Staab (2009). F-A Model of Events based on the Foundational Ontology DOLCE+DnSUltralite. In *K-CAP 2009*, pp. 137–144.

Segers, R. et al. (2011). Hacking history : Automatic historical event extraction for enriching cultural heritage multimedia collections. In *K-CAP 2011*.

Suchanek, F. M. et al. (2008). Yago : A large ontology from wikipedia and wordnet. *Web Semantics : Science, Services and Agents on the WWW 6*(3), 203–217.

TensorFlow (2016). TensorFlow. `https://www.tensorflow.org/`.

Tilley, A. (1992). *Plot snakes and the dynamics of narrative experience*. Univ. Press of Florida.

Van der Meij, L., A. Isaac, et C. Zinn (2010). A web-based repository service for vocabularies and alignments in the cultural heritage domain. In *ESWC 2010*, pp. 394–409.

Van Hage, W. et al. (2012). Abstracting and reasoning over ship trajectories and web data with the Simple Event Model (SEM). *Multimedia Tools and Applications 57*(1), 175–197.

Van Hooland, S. et al. (2015). Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities 30*(2), 262–279.

Yeung, C. et al. (2014). A knowledge extraction and representation system for narrative analysis in the construction industry. *Expert systems with applications 41*(13), 5710–5722.

Zarri, G. (2009). *Representation and management of narrative information : Theoretical principles and implementation*. Springer Science & Business Media.

Summary

An important aspect of cultural heritage preservation is the collection and collation of personal views and anecdotal stories of ordinary citizens. In this paper, we present a software architecture to facilitate narratives authoring and sharing about the time of the European construction in Luxembourg. More precisely, the proposed solution aims at supporting semi-automatic story input, where users can freely type their content, get automatic structure extraction as they type, along with related entities and information.