Modélisation et reformulation d'expressions temporelles extraites de textes en langage naturel

Cyril Faucher*, Jean-Yves Lafaye*, Frédéric Bertrand* et Charles Teissèdre**,***

*L3i, Université de La Rochelle, 17042 La Rochelle, France
{cyril.faucher, jean-yves.lafaye, frederic.bertrand}@univ-lr.fr,
**MoDyCo - Université de Paris Ouest Nanterre La Défense - CNRS, France
***Mondeca - 3, cité Nollez, Paris, France
charles.teissedre@mondeca.com,

Résumé. L'approche présentée modélise et reformule des expressions temporelles décrivant des événements périodiques extraites de textes en langage naturel tels que saisis par un humain. Ces informations temporelles permettent l'instanciation d'un modèle linguistique spécifique au domaine métier étudié. Pour développer une approche générique, nous avons spécifié un modèle objet pivot qui assure l'interopérabilité avec des standards comme iCalendar. Les expressions temporelles découvertes dans le texte et formalisées dans le modèle peuvent être modifiées *via* une interface graphique ou un éditeur de texte contrôlé. L'intérêt du mode texte est de produire une reformulation des informations saisies, dans un langage formel non ambigu, proche du langage naturel initial. Ceci est particulièrement adapté à des fins de validation sémantique par l'utilisateur. Notre approche se nomme « Temporal Knowledge Acquisition and Modeling » (TKAM) ; elle est implémentée sous forme d'une chaîne de traitements utilisant les techniques de l'Ingénierie Dirigée par les Modèles (IDM).

1 Introduction

L'acquisition de connaissances temporelles est un besoin pour nombre d'applications (gestion de périodes d'accès, planification de tâches, etc). Pour permettre cette acquisition il est nécessaire de disposer de métadonnées temporelles sur des phénomènes observables. Notre contribution concerne plus particulièrement les propriétés temporelles d'événements périodiques, mais gère également de façon plus classique les événements répétés ou ponctuels. Les interfaces graphiques de saisie de propriétés temporelles périodiques peuvent être complexes. Elles assurent une structuration de la saisie mais s'adaptent difficilement aux changements de contexte liés aux différents domaines d'application. La sémantique des interfaces, souvent figée, limite les possibilités d'évolution. Pour toutes ces raisons, il est utile d'assister l'utilisateur. La saisie de texte libre a l'avantage de s'adapter au contexte, l'utilisateur conserve son vocabulaire métier et ses habitudes linguistiques pour exprimer des propriétés temporelles. En revanche, de nombreuses imprécisions, ambiguïtés voire des contradictions peuvent être induites. Nous proposons dans cet article une chaîne de traitement (TKAM) permettant l'extraction d'informations à partir d'un texte initial, la modélisation des expressions temporelles

(modèles linguistique et pivot) et leur reformulation selon une grammaire formelle. Le modèle pivot permet d'interfacer à volonté des applications métier. La suite de l'article est organisée en trois sections. La Section 2 est consacrée à la description de la chaîne TKAM, puis la Section 3 se focalise sur un exemple d'acquisition de périodes d'accessibilité. Enfin la Section 4 conclut en présentant nos perspectives.

2 La chaîne d'acquisition de connaissances temporelles TKAM

La chaîne TKAM génère un texte contrôlé/structuré à partir d'un texte libre produit par un utilisateur. Un module de filtrage sélectionne les portions de texte contenant des informations temporelles. Puis un module d'annotation marque le texte suivant le type des expressions détectées. L'ensemble des annotations est ensuite utilisé pour instancier un modèle linguistique (Battistelli et al., 2008), (Teissèdre et al., 2010). Le modèle linguistique ne peut à ce stade qu'être spécifique et exploiter les connaissances métier particulières. Un exemple est donné dans la section suivante où le domaine concerne les périodes d'accessibilité de lieux publics (ex : « ouvert du lundi au vendredi, de 10h à 18h »).
Afin de rendre le système interopérable, nous utilisons un modèle pivot qui spécifie les occurrences temporelles en intension (ex. : « 3 fois par jour ») et (optionnellement) en extension (dates calendaires) (Faucher et al., 2010). La forme intensionnelle est concise et porte plus d'information que l'extension (i.e. : connaissance explicite des fréquences d'occurrence). Les dates calendaires peuvent être calculées à la volée. Ce modèle étend la norme ISO 19108 (ISO, 2002). Il possède notamment des concepts permettant de spécifier des fréquences (ex. : « 1 fois tous les 4 ans »), des intervalles récurrents (ex. : « tous les 1ers lundi de chaque mois de 14h à 16h ») et d'exprimer des positions relatives (relations d'Allen) entre des expressions (ex. : « 3 heures *avant* la basse mer »). Il est également possible de définir des exceptions notamment sous forme de règles périodiques. Le modèle pivot est doté d'une contrepartie textuelle permettant de spécifier n'importe quelle instance grâce à une grammaire formelle miroir du modèle pivot et proche du langage naturel. Le passage des instances du modèle linguistique à celles du modèle pivot est automatisé par des techniques de transformations (Kermeta [1]) relevant de l'Ingénierie Dirigée par les Modèles (Schmidt, 2006). Le passage des instances du modèle pivot aux phrases conformes à la grammaire utilise l'outil xText [2]. La définition d'une grammaire xText s'appuie explicitement sur les concepts du modèle. Le résultat est un texte contrôlé et interprétable par xText permettant de regénérer les instances originales. Les instances du modèle pivot permettent également d'instancier des modèles normalisés tels que iCalendar (Dawson et Stenerson, 1998).
Afin de contrôler d'éventuelles incohérences telles que des incompatibilités d'unités calendaires, des contraintes structurelles et sémantiques écrites en OCL [3] sont vérifiées lors de la transformation de modèles qui instancie le modèle pivot. Ces mêmes contraintes sont utilisées pour vérifier la cohérence des instances générées avec xText.

1. http://www.kermeta.org
2. http://www.eclipse.org/Xtext/
3. http://www.omg.org/spec/OCL/2.0/

3 Utilisation de la chaîne d'acquisition TKAW pour définir des périodes d'accessibilité

La chaîne TKAM a été appliquée aux propriétés temporelles concernant des périodes d'accessibilité (cf. Figure 1). Une période d'accessibilité correspond aux horaires d'ouverture (resp. fermeture) d'un lieu public tel qu'un cinéma, théâtre, musée, etc. L'objectif est de stocker des propriétés temporelles en intension (plutôt qu'en extension, ce qui serait coûteux) et de les interroger de la même façon (ex. : « Quels sont les restaurants ouverts chaque jour après 23h à Lyon ? »). Les informations temporelles définissant des périodes d'accessibilité sont généralement elles-mêmes exprimées en intension : « tous les jours sauf le dimanche ». Le modèle linguistique métier capture ces expressions temporelles en intension, une transformation de modèles (Kermeta) instancie ensuite le modèle pivot. A partir des instances du modèle pivot, un calendrier est généré pour afficher les dates et horaires d'ouverture. L'utilisateur peut interagir avec le calendrier pour modification ou mise à jour dans le respect des contraintes du modèle pivot. Un module de traduction (xText) reformule les instances du modèle du pivot, sous forme de texte en langage naturel contrôlé à des fins de validation (sémantique) des instances saisies. Ce module est bi-directionnel et peut instancier à l'inverse le modèle pivot à partir d'expressions temporelles saisies selon la grammaire formelle.
La Figure 1 montre des exemples de mise en correspondance des concepts provenant du modèle d'accessibilité avec ceux du modèle pivot. Ces mises en correspondance sont la base de l'implémentation de la transformation de modèles instanciant le modèle pivot. Par exemple une instance de *CalendarExpressionInterval* devient une instance de *PeriodicTimeInterval*.
La reformulation d'expressions a été expérimentée sur un corpus de 513 expressions fournies par un industriel dans le cadre du projet ANR RelaxMultiMedias 2 [4]. Ces expérimentations ont permis de détecter des incohérences dans les expressions comme des définitions d'intervalle incomplètes. Ainsi des expressions telles que : « Dimanche à 16h30 » ont été détectées comme incohérentes et ont pu être corrigées pour être de la forme : « Dimanche de 14h30 à 16h30 ».

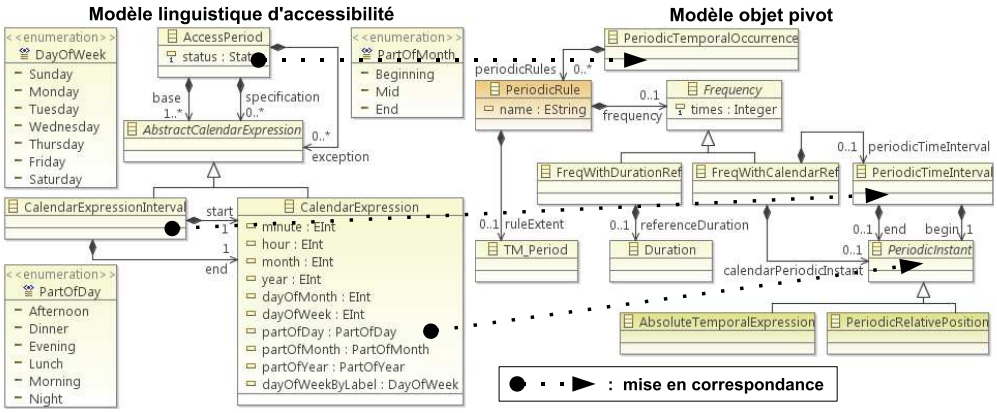

FIG. 1 – *Modèle de périodes d'accessibilité et modèle pivot.*

4. http://relaxmultimedia2.univ-lr.fr/

4 Conclusion et perspectives

L'approche présentée spécifie et implémente une chaîne complète de traitement de la langue naturelle pour extraire et gérer des expressions temporelles aussi bien en intension qu'en extension. Notre approche exploite les facilités de modélisation et de transformation de modèles propres à l'IDM. L'architecture proposée fait coopérer des modules de traitement de la langue naturelle, d'extraction de connaissance, d'instanciation d'un modèle pivot, d'IHM graphiques, de reformulation textuelle et de validation. Le modèle pivot proposé étend la norme ISO19108 pour la géométrie et la topologie des expressions temporelles et est compatible avec iCalendar pour la planification. Les travaux en cours portent sur l'extension de la sémantique des expressions temporelles prises en charge et sur les facultés de raisonnement (langages de règles, ontologies), tant sur les expressions calendaires que sur leurs représentations intensionnelles.

Remerciement Ce travail est financé par le projet ANR RelaxMultiMedias 2 (ContInt).

Références

Battistelli, D., J. Couto, J.-L. Minel, et S. R. Schwer (2008). Representing and visualizing calendar expressions in texts. In J. Bos et R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, Venice (Italy), pp. 365–373. College Publications.

Dawson, F. et D. Stenerson (1998). Internet calendaring and scheduling core object specification (icalendar) - rfc2445.

Faucher, C., C. Tissot, J.-Y. Lafaye, et F. Bertrand (2010). Benefits of a periodic temporal model for the simulation of human activities. In *GeoVA(t) (Geospatial Visual Analytics : Focus on Time)*, Guimaraes (Portugal).

ISO (2002). Text of 19108 geographic information - temporal schema.

Schmidt, D. C. (2006). Model-driven engineering. *IEEE Computer Society*.

Teissèdre, C., D. Battistelli, et J.-L. Minel (2010). Resources for calendar expressions semantic tagging and temporal navigation through texts. In *The seventh international conference on Language Resources and Evaluation (LREC)*, Valletta (Malta).

Summary

Our approach models and reformulates temporal expressions for periodic events. This temporal information is automatically extracted from natural language texts, and allows the instantiation of a domain specific linguistic model. Aiming at a generic approach, we specify a pivot object model which ensures the interoperability with such standards as iCalendar. The temporal expressions mined in the text and then stored in the pivot model, can be enriched and modified by a user *via* a graphical or textual controlled interface. The textual mode is invaluable since it retrieves an unambiguous formal language and keeps close to the user's natural language. This is well-fitted for a semantic validation of the data by the user. Our proposal is fully implemented through Model Driven Engineering (MDE) techniques.