Extracting information from French obituaries

Deryle W. Lonsdale
BYU Linguistics
Provo, UT
lonz@byu.edu

David W. Embley
BYU Computer Science
Provo, UT
embley@cs.byu.edu

Stephen W. Liddle
BYU Marriott School
Provo, UT
liddle@byu.edu

Joseph Park
BYU Computer Science
Provo, UT
jspark2012@gmail.com

Abstract

This paper discusses ongoing efforts to develop a system for extracting information from French obituaries. The approach is based on prior work on ontology-based English obituary extraction, which we first summarize. Then we mention sources for, and characteristics of, typical French obituaries. We sketch ongoing knowledge source development and the relevant tools, as well as an evaluation procedure we intend to follow for quantifying performance of the French system. Throughout we mention some lessons learned in the process.

1 Background

In previous work (Embley et al., 1999) we addressed extracting information from English obituaries. Our process involved the following steps:

1. Development of a narrow-domain ontology to model the concepts associated with an obituary.

2. Parsing the ontology to create a database schema and rules for matching textual information.

3. Processing web-based obituaries and recognizing/removing extralinguistic information (markup, record boundaries, etc.).

4. Marshalling recognizers that use the matching rules to extract textual information corresponding to the ontology's concepts.

5. Populating a relational database with the extracted results so that they can be queried via standard access methods.

The ontology was represented as an object-relationship model instance that specified concepts, relationships, and participation constraints (Embley et al., 1992). In the application in question they specified attributes that might be found in a person's obituary: the deceased's name, age, and death date; viewing, funeral, and interment information; relatives' names and their relationship with the deceased. Various types of knowledge sources were developed that assisted the system in locating and extracting this information. We processed over 130 English obituaries from two U.S. newspaper websites and evaluated the precision and recall of our results. Subsequently we also collected and processed another set of about 100 obituaries from across the English-speaking world.

In this paper we discuss subsequent related work: extracting similar information from French obituaries. Our goal is to apply the same (though updated) ontology-based extraction framework to processing a comparable corpus of obituaries written in French and appearing on the Web. We thus instantiate a French extraction system, while at the same time pushing the ontological framework and the extraction system beyond English-based processing to assure handling of more languages.

2 French obituaries

In our English obituary work we began by extracting information from obituaries printed in Utah and

Arizona newspapers. We were able to achieve, over a set of about 20 features (e.g. the deceased's age, birth date, death date, funeral address, viewing address, relatives' names) overall precision and recall rates of around 90%.

In order to investigate the robustness of our knowledge sources we also collected and analyzed obituaries from other parts of the English-speaking world: New Zealand, Sri Lanka, India, Ireland, and other U.S. states. As expected, different cultural conventions and dialectal differences in expressing obituary information resulted in interesting issues. For example, ceremonial events found elsewhere did not have corresponding U.S. concepts: the "tenth day kriya" in India or the "cortege" in Sri Lanka, for example. Consequently, precision and recall dropped slightly for these obituaries.

Overall French obituaries do not vary substantially from English ones: the facts reported are largely the same. They are found in places similar to where English ones are located: newspapers (both print and online versions), funeral home websites, and websites dedicated to the topic. For example, www.defunt.be is a Belgian French-language obituary site that lists and archives thousands of obituaries. Figure 1 shows a typical obituary from that repository.

French obituaries do tend to exhibit some variation throughout the French-speaking world, as mentioned earlier for English ones. Though to our knowledge no formal corpus analysis has been done on French obituaries and their content and variation, some observations are pertinent:

- European deceased are more often cremated, and their memorial services exhibit a wider variety of settings. European French obituaries also tend to mention more relatives and friends, though hardly ever a life history (education, careers, hobbies) of the deceased.

- French-Canadian obituaries and funerary procedures tend to more closely align with American ones.

- Swiss-French obituaries tend to contain more euphemisms for death, whereas the French and Belgian-French obituaries are less figurative and more direct.

Il a rejoint celle qui l'aimait, en nous laissant tout son amour.

Monsieur et Madame Guy SLUYSMANS - LAUMANS

son fils et sa belle-fille ;

Coraline et Laurent, Gaëtanne et Mike, Anaïs et Benoît, Guyllian et Mathilde, Océane

ses petits-enfants ;

Rayan, Kayla, Wyatt, la petite Princesse

ses arrière-petits-enfants ;

Ses soeurs, beaux-frères, neveux et nièces

Ses cousins et cousines

Les Familles SLUYSMANS - JONNIAUX et apparentées.

ont la grande tristesse de vous annoncer le décès de

veuf de Madame Marie JONNIAUX

Né à Limoges (France) le 23 juillet 1928 et décédé à Vezin le 16 décembre 2011

La liturgie des funérailles, suivie de l'inhumation au cimetière de Vezin se déroulera en l'Eglise Saint-Martin de Vezin le mardi 20 décembre 2011 à 10H30.

Le défunt repose au funérarium "Warzee" à Sclayn, rue Marche en Pré, 24.

Les visites ont lieu ces samedi, dimanche et lundi de 17 à 19 heures.

Le jour des funérailles, réunion à l'église.

Le présent avis tient lieu de faire-part

1970 Wezembeek-Oppem, rue Gergel, 106.
5300 Andenne (Vezin), rue des Priesses, 537

P.F et funérarium WARZEE
Andenne - Seilles - Sclayn - Petit-Warêt et régions.
085/82.31.35 - 085/84.50.71

Pompes Funèbres Warzée
Cliquez ici pour recevoir les infos décès de cette entreprise de Pompes Funèbres

Figure 1: Sample French obituary.

Some—though not all—obituary repositories offer some editorial control; a few have custom applications that guide writers through the authoring process, resulting in fairly formulaic, almost templatic obituaries.

3 Developing knowledge sources

Development of the extraction ontology for French obituaries was a relatively straightforward process. Our editor is an application that supports interactive encoding of the knowledge required for extraction of obituary information. Figure 2 displays a sample ontology. Though it doesn't include all information we will eventually want to extract from French obituaries, it is indicative of the types of attributes we want to target for extraction.

The primary object is the deceased person **Défunt**, and we assume that each person will be named explicitly **nomDéfunt**. Many European obituaries also include a title **titreDéfunt** for the deceased such as *Monsieur* (Mr.), *Madame* (Mrs.), *Mademoiselle* (Miss), or—as in the sample obituary

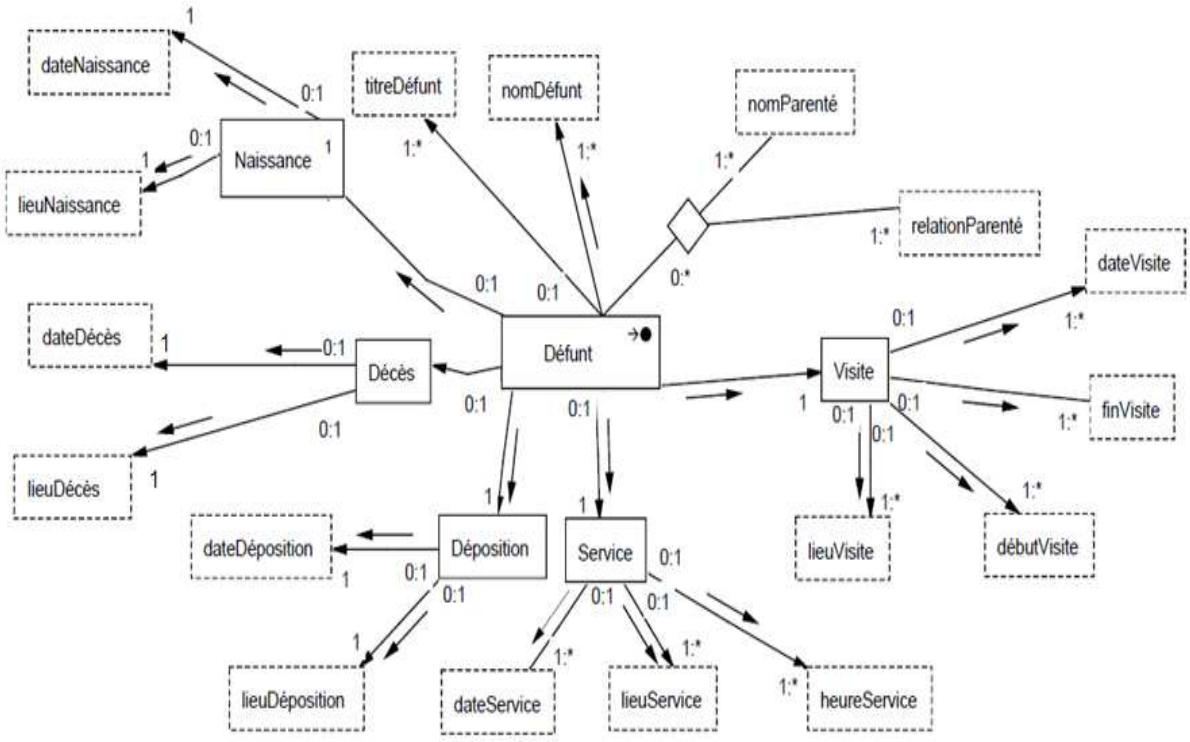

Figure 2: Sample French obituary extraction ontology.

above—*veuf* (widower), but that feature is specified as optional in the obituary.

Obituaries typically mention information about the person's birth **Naissance**: the birthdate **dateNaissance** and birthplace **lieuNaissance**. They also usually list information about the death **Décès**: the death date **lieuDécès** and death place **lieuDécès**.

There is usually some type of visit **Visite** to the deceased person's body arranged for interested parties; often in Europe the body lies at the deceased's home, though increasingly more often at a funeral home. Obituaries name the place **lieuVisite**, the date when visiting is possible **dateVisite**, along with the beginning **débutVisite** time and ending time **finVisite** for visits.

Some type of memorial service **Service** is also mentioned in most obituaries, along with its date **dateService**, time **heureService**, and location **lieuService**. Final disposition of the body **Déposition** is also announced, as well as the place **lieuDéposition**.

Finally, the ontology specifies a relationship set to account for any relatives mentioned: each relative's name **nomParenté** is paired with a documented relationship **relationParenté**.

Of course, other information often found in obituaries could also be included in the ontology and targeted for extraction. For example, many French obituaries express thanks and acknowledgement to caregivers, medical personnel, hospitals, and hospices. Many also mention organizations or accounts to which donations may be directed in memory of the deceased. The type of disposition of the body (i.e. burial, cremation, inurnment) is also often mentioned. Deciding which of these items of information is of long-term value determines what should be extracted, and hence what needs to be included in the ontology.

3.1 Lexical knowledge sources

Building the lexicons for was relatively straightforward. For example, lists of first names and last names were readily available at French womens' magazines baby naming websites. Some 280,000

family names and a few thousand given names[1] were gathered in this way.

At the current time we have completed a first version of the French obituary extraction ontology and the associated knowledge sources. For straightforward matches the system performs well. The French obituary files we have collected are encoded in a wide variety of character representation formats, but using the inconv tool we were able to convert them to UTF-8 format, which our extraction tool supports. Some of the user interface functions—for example, the ontology editor—require basic internationalization updates to support working with French data.

One problem that we encountered in using our system is that in Europe peoples' last names are often represented entirely in uppercase. Coupled with the fact that uppercase letters are not typically accented in Europe, we had to create capitalized, unaccented versions of the family names we located, doubling the size of the last name lexicon to over a half million entries. In the future we will add a method to address this issue programmatically.

Place name lists were easily converted from online sources for gazetteers, resulting in tens of thousands of city, town, and village names in French-speaking Europe and Canada.

Since the narrative in European obituaries is relatively terse, anaphor is frequent, which will prove problematic for our extraction. For example, a typical pattern encountered is:

né à Paris et y décédé
born in Paris and died there

where the pronoun "y" is an anaphor for the prepositional phrase "in Paris".

4 Evaluation and discussion

As the extraction system finds matches in the obituary text against the ontology, the matches and their locations are stored in an XML file for subsequent processing such as querying and results table generation.

We have harvested a corpus of 2,000 randomly selected representative French obituaries, creating from them a training/testing partition (75%/25%). We are currently preparing to annotate the testing partition to create a gold standard for subsequent evaluation. A tool we have designed for data annotation allows a person to view an obituary and indicate relevant items: the deceased's name, the death date, etc. with minimal keystroke and mouse operations. The items' offset and type information is stored for later use.

We have also written a metric calculation tool that takes an extraction results XML file and its corresponding hand-annotated gold-standard version. The tool compares the two, computes precision and recall measures from the extraction against the expected matches, and reports results.

At press time we have not yet run a comprehensive end-to-end evaluation of the French ontology extraction, but we have successfully run individual components in the processing pipeline. A complete evaluation will take place in the near future.

[1] Until relatively recently given names were strictly controlled by the government in France.

References

Embley, D., Campbell, D., Jiang, Y., Liddle, S., Lonsdale, D., Ng, Y.-K., and Smith, R. (1999). Conceptual-model-based data extraction from multiple-record web pages. *Data and Knowledge Engineering*, 31(3):227–251.

Embley, D. W., Kurtz, B., and Woodfield, S. (1992). *Object-oriented Systems Analysis: A Model-Driven Approach*. Prentice Hall.