Enrichissement de données en breton avec Wordnet

Annie Foret
IRISA & Université Rennes 1, 35042 Rennes Cedex, France
foret@irisa.fr

RÉSUMÉ

Nous décrivons une expérience d'enrichissement automatique de données en breton. Les données sont des unités de texte en breton. Certaines unités sont enrichies avec des synsets (synonym sets) de Wordnets en exploitant d'une part les ressources d'Apertium pour la paire de langues breton et français et d'autre part des ressources de type Wordnet pour le français et pour l'anglais. Le résultat peut-être visualisé et exploré de diverses manières : notre réalisation est sous forme de système d'information interactif. Notre approche repose d'une part sur des chaînes automatiques de traitements linguistiques en breton et en français et sur un environnement d'exploration de systèmes d'information logiques.

ABSTRACT

Breton data enrichment with Wordnet.

We describe an automatic data enrichment experiment in breton. The data consists in text units in breton. Some units are enriched with synsets (synonym sets) of wordnets exploiting Apertium resources for the language pair breton and french and Wordnet resources for french and english. The result can be viewed and explored in various ways : our proposal is in the form of an interactive information system. Our approach is based on an automatic tool chain for natural language processing in breton and french and a platform for logical information systems.

MOTS-CLÉS : breton, lexique, wordnet, système d'information, recherche d'information, sémantique.

KEYWORDS: breton, lexicon, wordnet, information system, information retrieval, semantics.

1 Introduction

Nous décrivons une expérience d'enrichissement automatique de données en breton et nous présentons la réalisation en cours de cette chaîne de traitement. Les données initiales sont des unités de texte en breton (Foret *et al.*, 2015) saisies manuellement à partir d'un ouvrage d'une série illustrée Les "Mille premiers mots en breton" (Kergoat *et al.*, 2007). Ce lexique, bien que restreint (la chaîne de traitement pourrait s'appliquer à des lexiques plus étendus), offre un vocabulaire de base, de référence pour le breton, et utile aux apprenants. Il a de plus été saisi en conservant l'organisation thématique du livre, et en ajoutant des indications spécifiques au breton : les *mutations* en breton sont les variations de la consonne initiale, (par exemple, l'expression *an daol* pour "une table" est indiqué en figure 1 par : *an d|taol*, le lemme du nom étant *taol*, et la mutation avec cet article *an* étant $d > t$) une description est fournie sur Le site ARBRES (`http://arbres.iker.cnrs.fr`) (Jouitteau, 2005) et une approche pour les gérer dans (Poibeau, 2014).

Certaines unités sont enrichies avec des groupes de synonymes, synsets (synonym sets) de Wordnet en exploitant d'une part les ressources GPL d'Apertium (Tyers, 2010) `apertium.org` pour la paire de langues breton et français et d'autre part des ressources de type Wordnet pour le français et pour l'anglais. Le résultat peut-être visualisé et exploré de diverses manières : notre réalisation est sous forme de système d'information interactif. Notre approche repose d'une part sur des chaînes automatiques de traitements linguistiques en breton et en français et sur un environnement d'exploration de données basé sur les systèmes d'information logiques.

Une chaîne de traitement construisant un système d'information lexical à explorer à partir d'articles en français et en anglais a été proposé par (Cellier *et al.*, 2016). Notre objectif général est assez proche et vise un système d'information interactif, utile et d'emploi sûr et aisé ; mais la chaîne de traitement présentée ici concerne le breton, pour lequel certains problèmes et traitements sont bien sûr spécifiques. Une autre caractéristique de cette réalisation est la possibilité de l'utiliser localement (hors connexion internet) [1].

2 Jeu de donnnées et présentation initiale

Nous considérons pour cette réalisation, un lexique saisi (Foret *et al.*, 2015) à partir du livre les "Mille premiers mots en breton" (Kergoat *et al.*, 2007).

Dans (Foret *et al.*, 2015), le lexique est ensuite chargé comme système d'information avec des facettes logiques, dans l'outil Camelis (version 1, accessible à `http://www.irisa.fr/LIS/ferre/camelis/`) : Camelis (Ferré, 2009; Ferré & Ridoux, 2004; Ferré & Ridoux, 2004) est basé sur une extension de l'analyse de concepts formels (Ganter & Wille, 1999) et peut gérer des hiérarchies de propriétés assez générales, il s'agit en ce sens d'un système de gestion de contextes et de propriétés logiques ; les couples propriétés et objets les vérifiant forment un *treillis de concepts* comme une facette spécifique fermée ou ouverte selon le type et le niveau d'exploration et de filtrage/sélection choisis.

Rôles des fenêtres Camelis par rapport à un contexte. L'outil Camelis, chargé avec un contexte initial, présente trois fenêtres relatives à un contexte courant, qui évolue au fil des sélections dans ces fenêtres. Un tel contexte peut être hétérogène, il peut contenir plus de sortes d'objets et de propriétés, selon les préférences et les usages prévus.

Fenêtre d'objets : la partie droite présente les objets du contexte courant, par leur label.

Fenêtre de propriétés : la partie gauche indique les propriétés, organisées en arbres selon les relations entre les propriétés. Il s'agit d'un index cliquable qui permet de passer d'un contexte à un autre. Les cardinalités des liens/sous-contextes y sont aussi affichées.

Fenêtre de requête : la partie du haut contient une requête caractérisant le contexte courant : c'est une propriété satisfaite par tous les objets du contexte courant ; elle n'a pas besoin d'être saisie puisqu'elle est mise à jour automatiquement selon les sélections dans les deux autres fenêtres. L'utilisation ne nécessite pas de connaissance *a priori*, mais il est aussi possible de rédiger directement les requêtes.

1. y compris les synsets Wordnet, mais hormis les liens-action vers Babelnet via http

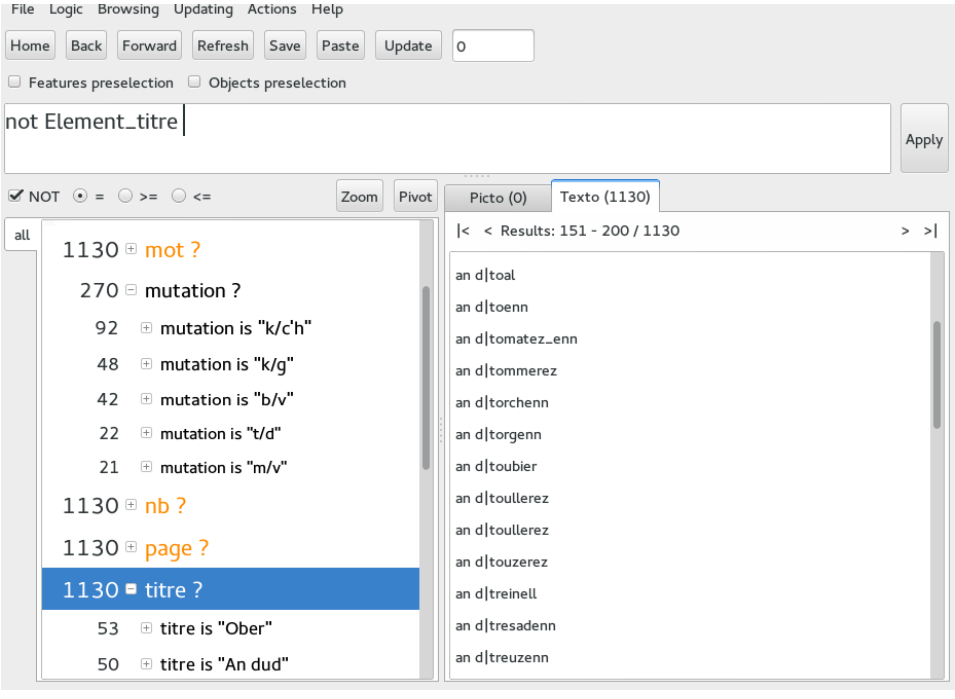

FIGURE 1 – Capture d'écran de Camelis, avec le contexte de départ

Pour ce petit lexique plusieurs sortes d'objets sont distinguées :

— des titres représentant des thèmes comme ("à la maison", "à l'école", etc.) ; ceux-ci pouvent comporter des variantes (c'est le cas de "Ar mezeg" pour le titre is "Ar medisin", désignant "le médecin"), voir la figure 2 ;

— et les expressions rattachées à un titre (correspondant à une page ou un intervalle de pages dans le livre) ; la figure 1 présente un contexte courant, avec à droite l'objet "an dltaol" suivi d'autres mots subissant la même mutation, la fenêtre en haut affiche la propriété sélectionnée (les objets courants ne sont pas des titres), et la fenêtre à gauche est une vue de l'arbre courant des propriétés (cliquable pour affiner la recherche).

Les objets sont étiquetés par leur classe, cette information peut être structurée comme ici en hiérarchie taxonomique et affichée par Camelis dans l'index de navigation à gauche, comme dans la figure 4.

Remarque. En poursuivant cet exemple de terme, on peut noter que Babelnet (utilisé plus loin) propose le breton dans sa liste de langages, mais ne reconnaît pas correctement le mot "mezeg", désignant "médecin", pourtant considéré dans un vocabulaire de base ; c'est ce que montre ce simple test : `http://babelnet.org/search?word=mezeg&lang=BR` qui donne un résultat, mais pour un nom de lieu.

3 Enrichissement avec les synsets de wordnet

Wordnet (wordnet.princeton.edu) est un réseau lexical d'abord développé pour l'anglais, qui sert aussi de référence pour d'autres langues, où des codes *synset* regroupent des ensembles de synonymes. Pour l'anglais, une forme XML est proposée par (Lapalme, 2014). Pour le français, nous avons exploité une autre ressource XML, appelée WoNef, accessible à wonef.fr et qui permet de relier les unités de sens dans les deux langues (par les codes synset).

Réalisation. Le contexte produit présente les unités de sens par leurs codes et ensemble de mots en français. Nous avons privilégié ici les mots en français, mais nous pourrions procéder de même pour associer la liste de mots en anglais, à partir de la version XML de Wordnet.

Cette construction utilise la paire br-fr de Apertium. Pour chaque langue ajoutée, les mots inconnus d'Apertium sont marqués par * (au début). Notons que la traduction d'une forme dictionnaire pour un terme du lexique breton n'est cependant pas toujours un lemme dans la langue cible : tel que le pluriel pour un nom collectif en breton sans suffixe visible [2], par exemple "ar *gwez*" pour "les *arbres*".

Actuellement, pour l'étape d'ajout des synsets comme propriété, nous considérons uniquement les expressions du lexique AvecArticle (cela sélectionne la plupart des termes du lexique : 1000 mots, dont des titres), ce qui amène à relier dans le sous-ensemble des noms (code n) de Wordnet. Cependant au stade actuel, l'association est partielle (pour 416 unités, avec en moyenne 7,5 synsets par unité) ; les termes sans synset en propriété comprennent notamment ceux sans traduction par Apertium br-fr (voir (Foret *et al.*, 2015)).

Remarques. Une première méthode consisterait à appliquer TreeTagger sur le mot français afin de relier ensuite le terme selon le lemme du mot français. Cette piste a été amorcée mais non poursuivie, il faudrait disposer auparavant de la catégorie pour améliorer les résultats. Une alternative est d'exploiter l'ensemble des outils de Apertium pour produire plus de détails dans chacune des deux langues. Une autre difficulté concerne les expressions composées.

4 Actions associées

Lien à Babelnet. Babelnet (https://en.wikipedia.org/wiki/BabelNet) est un réseau sémantique multilingue à très large couverture. Il a été construit automatiquement, notamment avec l'encyclopédie Wikipedia et Wordnet. Comme dans Wordnet les mots sont regroupés en ensembles de synonymes : les *label synsets*. En pratique, les codes synsets de Wordnet peuvent être utilisés en les préfixant par wn:, c'est un procédé que nous pourrions utiliser ici.

Nous avons actuellement réalisé le lien à travers la traduction en français, en deux points :

- dans le fichier d'information principal, où chaque objet est écrit par une ligne (voir figure 4), nous ajoutons *sur chaque ligne*, que le mot en français est un argument possible (pour une commande interactive), par exemple avec : {"fr", "cmdFR", "Couverture"} pour le nom "ar golo"

cet objet appartient dans le contexte au thème du voyage, de titre "beajiñ", voir figure 4 ;

2. voir http://arbres.iker.cnrs.fr/index.php?title=Noms_collectifs

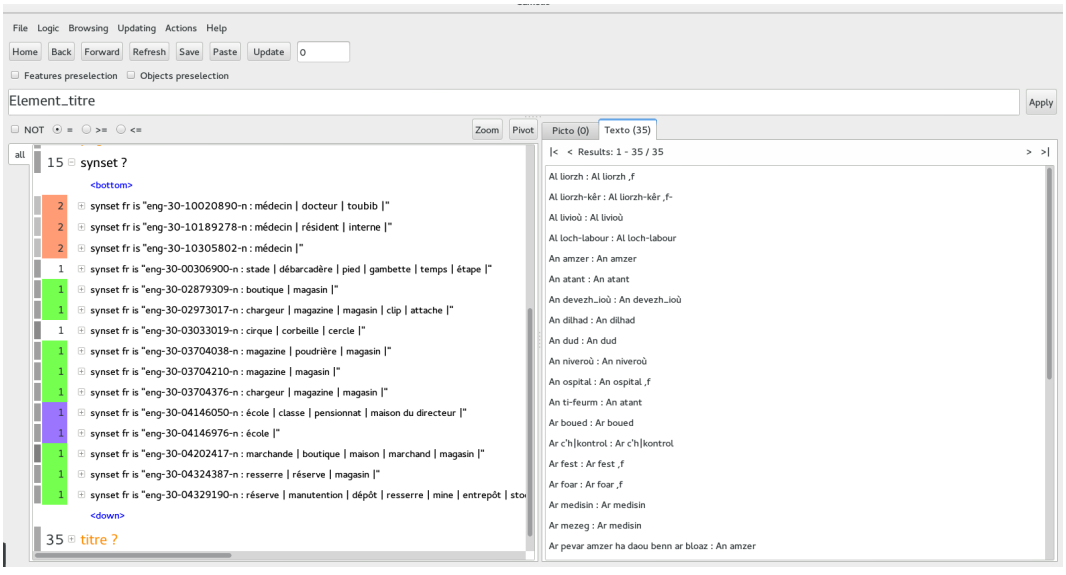

FIGURE 2 – Capture d'écran de Camelis, montrant les synsets obtenus pour des titres

```
mk {"id", "cmdLabel", "ar golo : Beajiñ"} {"titre", "cmdTopic", "Couverture"}
{"ref", "cmdExpr", "ar golo"} {"fr", "cmdFR", "Couverture"}
 "ar golo : Beajiñ" nb Apertium = 2,mot gb is "*Litkovrilo",mot fr is "Couverture",mot eo is "Litkovrilo",
 synset fr is "eng-30-06389398-n : fourchette | couverture |",
 synset fr is "eng-30-04605726-n : papier d'emballage | couverture | emballage | peignoir | papier |",
 synset fr is "eng-30-04605446-n : couverture | peignoir |",
 synset fr is "eng-30-04118021-n : moquette | tapis | plaid | couverture | petit tapis | carpette | descente de lit | postiche | moumoute
 synset fr is "eng-30-02849154-n : nappe | couverture | couvrante |",
 synset fr is "eng-30-01049685-n : masquage | couverture | recouvrement | plaque | enveloppe |",
...
MilleMotsBzh,mot Ref is "ar golo",mot Dico is "golo",AvecArticle
titre is "Beajiñ",Element_mot,page in [20,21]
```

FIGURE 3 – Extrait du fichier de contexte, pour le terme "ar golo" du thème voyage

- dans le fichier d'information, nous indiquons par *une ligne générique* le choix d'action, cette ligne comprend quatre parties, le mot action, le nom de l'action, la commande avec l'argument `$(fr)` et la propriété de filtrage de contexte (ici `all` pour l'ensemble) :

```
action "cmdFR" "firefox \"http://babelnet.org/search?word=$(fr)&lang=FR\"  & " all
```

En suivant le même principe et selon les préférences, d'autres actions peut être prévues, par exemple pour éditer ou interroger localement (par un outil XPATH, comme BaseX) le fichier de ressources XML Wonef (ou pour l'anglais Wordnet en version XML).

Ainsi, à une étape de navigation, en cliquant dans la fenêtre des objets, l'utilisateur verra les actions associées à un objet du contexte courant et pourra en déclencher.

5 Conclusions et perspectives

Nous avons proposé une chaîne de traitements qui enrichit des données en breton, avec des informations d'un réseau sémantique de type Wordnet. Les données enrichies peuvent alors être chargées

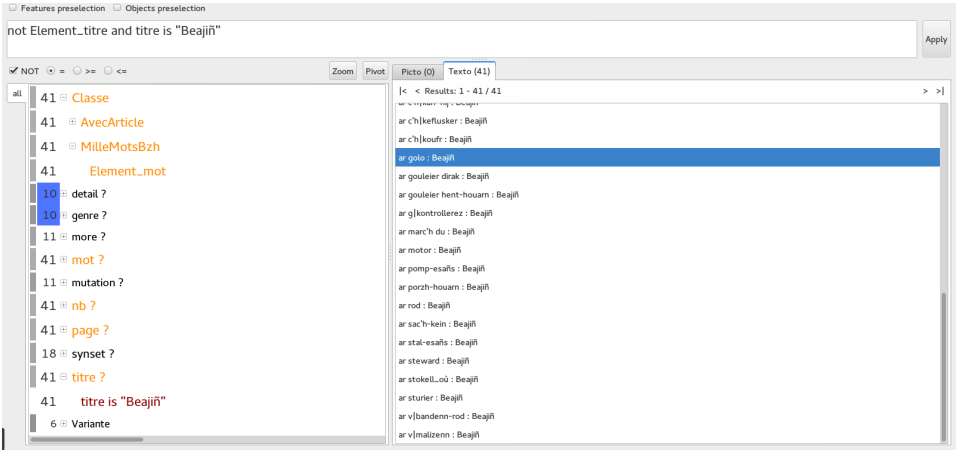

FIGURE 4 – Capture d'écran de Camelis, montrant les mots du thème voyage

comme système d'information (loqique) et être explorées de diverses façons : en multi-facettes par simples sélections successives dans un arbre de propriétés avec des liens vers d'autres ressources. Ce système pourra être proposé à un apprenant du breton pour faciliter sa recherche de l'information. Cela peut aussi servir au spécialiste, par exemple pour une forme d'évaluation d'une ressource par rapport à une autre (Foret *et al.*, 2015).

L'outil Camelis utilisé dans cette étude permet une exploration sûre (jamais de réponse vide, tout ce qui est accessible l'est en suivant l'arbre de navigation) généralisant en particulier les interrogations de type hiérarchique, et bases de données avec des outils de l'analyse de concepts logiques.

Un aspect important de ce travail avec des étapes automatisées est sa réutilisabilité : pour de nouvelles versions ; pour d'autres données à caractéristiques proches.

Il s'agit d'un travail en cours, [3] la couverture étant encore partielle en ce qui concerne l'association directe aux objets des synsets Wordnet (avec leur ensemble de mots en français).

Mise à jour. L'outil de gestion de contexte permet non seulement une navigation flexible, il est aussi prévu pour permettre la modification interactivement. Nous avons présenté ici un lien automatisé avec un code Wordnet (code français et anglais compatibles), mais ce type de lien pourrait être aussi repris manuellement, depuis l'outil Camelis interactif puis exporté en nouveau fichier de contexte (c'est aussi un fichier texte facilement modifiable, avec un objet décrit par ligne).

Web sémantique. Une version orientée *web sémantique* (Hitzler *et al.*, 2009) est une variante possible de ce travail. D'une part les ressources Wordnet et Babelnet ont des versions et des liens adaptés à SPARQL (`http://babelnet.org/sparql/`) D'autre part, les ressources dans ce format (RDF, ou un équivalent) peuvent être explorées avec l'outil Sparklis (Ferré, 2014), à la place du système Camelis.

3. une mise à disposition est prévue ici, en licence compatible GPL : `http://www.irisa.fr/LIS/softwares-fr`

Références

CELLIER P., FERRÉ S., FORET A. & RIDOUX O. (2016). Exploration des données du défi EGC 2016 à l'aide d'un système d'information logique. In C. DE RUNZ & B. CRÉMILLEUX, Eds., *16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016, 18-22 Janvier 2016, Reims, France*, volume E-30 of *RNTI*, p. 443–448 : Hermann-Éditions.

FERRÉ S. (2009). Camelis : a logical information system to organize and browse a collection of documents. *Int. J. General Systems*, **38**(4).

FERRÉ S. (2014). Expressive and scalable query-based faceted search over SPARQL endpoints. In P. MIKA & T. TUDORACHE, Eds., *Int. Semantic Web Conf.* : Springer.

FERRÉ S. & RIDOUX O. (2004). Introduction to logical information systems. *Inf. Process. Manage.*, **40**(3), 383–419.

FERRÉ S. & RIDOUX O. (2004). An introduction to logical information systems. *Information Processing & Management*, **40**(3), 383–419.

FORET A., BELLYNCK V. & BOITET C. (2015). Akenou-breizh, un projet de plate-forme valorisant des ressources et outils informatiques et linguistiques pour le breton. In *Actes de la Traitement Automatique des Langues Régionales de France et d'Europe*, Caen, France : Association pour le Traitement Automatique des Langues.

GANTER B. & WILLE R. (1999). *Formal Concept Analysis - Mathematical Foundations*. Springer.

HITZLER P., KRÖTZSCH M. & RUDOLPH S. (2009). *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC.

JOUITTEAU M. (2005). *La syntaxe comparée du breton, une enquête sur la périphérie gauche de la phrase bretonne*. PhD thesis, Nantes, France.

KERGOAT L., AMERY H. & CARTWRIGHT S. (2007). *Les 1000 premiers mots en breton*. Skol an Emsav, 8 edition.

LAPALME G. (2014). Wordnet en XML-HTML. In *TALN 2014 - Atelier RLTLN*.

POIBEAU T. (2014). Processing mutations in breton with finite-state transducers. In *Proceedings of the First Celtic Language Technology Workshop*, p. 28–32, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

TYERS F. M. (2010). Rule-based breton to french machine translation. In *Proceedings of the 14th Annual Conference of the European Association of Machine Translation, EAMT10*, p. 174–181.