Normalisation ISO/AFNOR des métadonnées (ISO TC37)

Gil Francopoulo / Tagmatica

www.tagmatica.com

1 Contexte

Le présent texte porte sur le processus d'annotation de vastes corpus de textes.

L'annotation ne peut plus être réalisée sans avoir recours aux techniques conjuguées de traitement automatique du langage et de représentation d'ontologies. En effet, la taille des corpus atteint maintenant des centaines de millions de mots, et de plus, bien souvent, ce ne sont pas des corpus figés, mais au contraire des textes en perpétuelle transformation. Ce qui exclut l'annotation manuelle de manière irrémédiable.

2 La solution

La seule solution consiste à travailler à deux niveaux :

- l'être humain intervient afin de déclarer des métadonnées et des règles de repérage de ces métadonnées ;
- le second niveau est réalisé par des programmes qui sont paramétrés par le premier niveau.

Mais ce n'est pas suffisant, il faut que le résultat, c'est-à-dire l'annotation, soit définie précisément et que sa spécification soit connue de tous les acteurs techniques. Pour ce faire, le canal de la normalisation ISO semble être un outil de communication bien adapté.

3 Les standards

La question des standards dans le domaine du TAL apparaît comme cruciale. Tout traitement linguistique requiert des ressources linguistiques (ex: lexiques) et met en jeu des outils de diverses natures.

L'absence de formats standardisés pour ces ressources et pour les entrées/sorties des outils nuit au déploiement des chaînes de traitement génériques, où il serait facile :

- d'ajouter/remplacer des composants,
- de fusionner les lexiques et des annotations.

Les efforts de standardisation sont concentrés au sein de l'ISO TC37 qui coordonne, au travers d'experts du monde entier, les propositions techniques pour l'annotation et la représentation des lexiques.

4 Les réquisits

Le problème n'est pas simple.

Intuitivement, on peut être à peu près certain que :

- les informations vont dépendre de la langue traitée,

- les présupposés théoriques seront différents selon les linguistes,
- les genres et qualités des textes ne vont pas autoriser les mêmes traitements.

De plus, les spécifications en question doivent obtenir un consensus technique au sein de l'ISO.

5 La méthode

La méthode choisie consiste à séparer le problème en deux : en distinguant la structure de la décoration.

Concernant la structure, la réponse de l'ISO-TC37 consiste en cinq standards majeurs sur l'annotation, les lexiques et les thésaurus.

- Morpho-syntactic Annotation Framework (MAF ISO 24611)
- Syntactic Annotation Framework (SynAF ISO 24615)
- Semantic Annotation Framework (SemAF ISO 24617)
- Lexical Markup Framework (LMF ISO 24613)
- Terminological Markup Framework (TMF ISO 16642)

Concernant la décoration, les métadonnées sont enregistrées de manière consensuelle dans le registre de catégories de données de l'ISO dans le cadre de la révision de l'ISO 12620. Le registre est en quelque sorte un « réservoir » dans lequel l'utilisateur va « piocher » des données, soit pour nommer un attribut, soit pour valuer un attribut par une constante. On utilise aussi les autres standards préexistants comme les noms de langues ou les noms des scripts.

Afin d'offrir une bonne souplesse et une grande puissance d'expression, l'association entre la structure et la décoration est très flexible.

Voyons sur un exemple concret de dictionnaire conforme à LMF.

```
<LexicalResource dtdVersion="16">
<GlobalInformation
   <feat att="languageCoding" val="ISO 639-3"/>
</GlobalInformation>
<Lexicon>
   <feat att="language" val="fra"/>
   <LexicalEntry morphologicalPatterns="AsPassif">
      <feat att="partOfSpeech" val="adjective"/>
      <Lemma>
          <feat att="writtenForm" val="actif"/>
      </Lemma>
etc.
```

Les éléments "LexicalResource", "GlobalInformation", "Lexicon", "LexicalEntry", "Lemma" sont des éléments de structure. Au contraire, les informations apparaissant en tant que traits comme "languageCoding", "language", "partOfSpeech", "adjective", "writtenForm" font partie de la décoration. Ils ont été pris dans le registre de catégories de données. Notons que les informations "ISO 639-3" font référence à un nom de norme, la valeur "fra" fait référence à la valeur du registre des langues et que la chaîne de caractères "actif" est la graphie du mot.

6 Le travail de description dans le registre

Le registre est découpé en profiles, chacun étant associé à un groupe d'experts. Les registres portent sur des domaines comme la morpho-syntaxe, la syntaxe, la sémantique lexicales etc. Les groupes de travail opèrent essentiellement à distance via le logiciel de gestion de catégories de données qui est hébergé à Nancy à l'adresse "http://syntax.inist.fr".

Les valeurs viennent de listes préexistantes comme:

- EAGLES pour les langues de l'Europe de l'ouest,
- MULTEXT-East pour les langues de l'Europe de l'est,
- une liste venant de l'Université de Sfax pour les langues sémitiques.

De plus, quelques valeurs issues des travaux MAF, SynAF et LMF ont été ajoutées. Notons par ailleurs qu'un travail en cours dans le projet NEDO consiste à établir une liste de valeurs pour les langues asiatiques. Une liste similaire est en cours de constitution pour les langues africaines par la délégation de l'Afrique du Sud. Lorsqu'elles seront stabilisés ces listes seront fusionnées avec le contenu actuel du registre.

7 Conclusion

Le registre est loin d'être complet mais il commence à être utilisé au sein de différentes implémentations fondées sur les spécifications de l'ISO-TC37. L'idée est de proposer une liste initiale qui sera progressivement complétée suite aux retours d'expériences. L'ambition est que le registre devienne à terme le point de référence des lexiques et annotations dans le contexte du traitement automatique du langage.

Bibliographie :

Francopoulo G., Declerck T., Sornlertlanvanich V., De la Clergerie E., Monachini M. 2008 Data Category Registry: morpho-syntactic and syntactic profiles. workshop: use and usage of language resource-related standards / LREC Marrakech