

Laboratoire
d'Informatique
de Robotique
et de Microélectronique
de Montpellier

Utilisation de techniques de TAL pour le mapping de schemas

Mathieu Roche

*Equipe TAL, LIRMM,
Université Montpellier 2*

**FORUM,
mars 2006**



- **Techniques terminologiques pour le mapping de schemas**
 - ➔ **1^{ère} étape** : Normalisation des labels
 - ➔ **2^{ème} étape** : Mapping des labels
 - Comparaison de préfixe/suffixe,
 - Mesures de similarité (exemple, Edit distance et String Matching),
 - n-grammes.
- **Perspectives dans le cadre de FORUM**
 - ➔ Techniques hybrides (terminologiques et contextuelles)

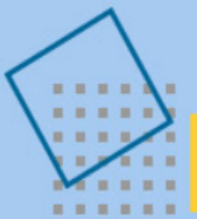


Techniques terminologiques pour le mapping de schemas

LIRMM

Techniques terminologiques

- Difficultés d'utiliser des techniques de TAL (Traitement Automatique du Langage) pour le mapping de schémas:
 - absence d'informations syntaxiques,
 - lexiques spécialisés utilisés (*par exemple, NomAuteur = nom d'un auteur*),
 - *etc.*
- Cet exposé s'appuie en grande partie sur l'état de l'art de la thèse de **Hassen Kefi** (LRI).

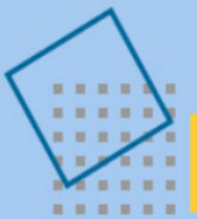


Normalisation des labels (1/3)

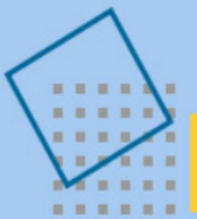
LIRMM

Techniques terminologiques > normalisation des labels

- Normalisation de labels nécessaire (avec souvent une intervention de l'expert, particulièrement dans le cas de données spécialisées telles que celles de FORUM !).
- **Types de normalisations** pouvant être effectuées :
 - **Segmentation des termes** des labels suivant les signes de ponctuations, les mots en majuscules, les symboles spéciaux.
✓ *Exemple : NomAuteur -> Nom Auteur*
 - **Elimination des “mots outils”** (articles, conjonctions, prépositions). *Question : Est-ce une solution adaptée ? (particulièrement pour les locutions figées telle que “pomme de terre”).*



- **Types de normalisations (suite) :**
 - **Expansion des abréviations terminologiques** (par exemple, les acronymes).
 - ✓ *Exemple* : TAL = traitement automatique du langage
 - Difficultés :
 - Mise à jour nécessaire des lexiques utilisés (*problème pour les acronymes récents tels que CNE*).
 - Choix de l'acronyme adapté (*CNE signifie-t-il : Caisse Nationale d'Épargne, Compagnie Nationale des Experts, Comité National d'Évaluation, Conseil National de l'Emballage, Contrat Nouvelle Embauche ?*)
 - Connaître la définition des acronymes des domaines spécialisés tels que ceux de FORUM.



Normalisation des labels (3/3)

LIRMM

Techniques terminologiques > normalisation des labels

- **Types de normalisations (suite) :**
 - **Lemmatisation** : déterminer la forme normalisée des mots (masculin pour les noms, infinitif pour les verbes, masculin singulier pour les adjectifs, etc).
 - ✓ *Exemple* : grèves étudiantes -> grève étudiant



- **Techniques terminologiques pour le mapping de schemas**
 - **1^{ère} étape** : Normalisation des labels
 - **2^{ème} étape** : Mapping des labels
 - Comparaison de préfixe/suffixe,
 - Mesures de similarité (exemple, Edit distance et String Matching),
 - n-grammes.
- **Perspectives dans le cadre de FORUM**
 - Techniques hybrides (terminologiques et contextuelles)

- But : vérifier qu'une chaîne de caractères $Ch1$ se retrouve :
 - au début d'une chaîne de caractères $Ch2$ (**préfixe**),
 - à la fin d'une chaîne de caractères $Ch2$ (**suffixe**).
- ✓ *Exemples de similarités :*
 - Préfixe -> $Ch1 = \text{chat}$ / $Ch2 = \text{chaton}$
 - Suffixe -> $Ch1 = \text{suivre}$ / $Ch2 = \text{poursuivre}$



Mesures de similarité ^(1/2)

LIRMM

Techniques terminologiques > Mapping de labels > Mesure de similarité

- Il existe de nombreuses mesures de similarité (pas seulement dans la littérature du “matching de schémas”).
- Exemple avec la distance « Edit distance » (notée E) = somme minimale du coût des opérations qu'il faut effectuer pour transformer $Ch1$ en $Ch2$.
Opérations : suppression, insertion, remplacement

- ✓ Exemple : $E(\text{gréviste}, \text{grève}) = 4$

Ch1 :	g	r	é	v	i	s	t	e
Opérations :			Remplacement		Insertion	Insertion	Insertion	
Ch2 :	g	r	è	v				e

- Mesure prenant en compte E : la mesure String Matching (SM) de Maedche et Staab :

$$SM(Ch1, Ch2) =$$

$$\max[0; (\min(|Ch1|, |Ch2|) - E(Ch1, Ch2)) / \min(|Ch1|, |Ch2|)]$$

- ✓ $SM(\text{gréviste}, \text{grève}) = \max(0; (5-4)/5) = 0.2$



Les n-grammes

LIRMM

Techniques terminologiques > Mapping de labels > n-grammes

- Technique des n -grammes est utilisée pour calculer le nombre de n caractères consécutifs.
- Généralement, la valeur de n varie entre 1 et 5.
 - ✓ *Exemple de tri-grammes : Ch1 = chat / Ch2 = chaton :*
 - $tri_gramme(Ch1) = \{cha, hat\}$
 - $tri_gramme(Ch2) = \{cha, hat, ato, ton\}$
- *Mise en oeuvre de mesures fondées sur les tri-grammes tels que la mesure de Lin.*



- **Techniques terminologiques pour le mapping de schemas**
 - **1^{ère} étape** : Normalisation des labels
 - **2^{ème} étape** : Mapping des labels
 - Comparaison de préfixe/suffixe,
 - Mesures de similarité (exemple, Edit distance et String Matching),
 - n-grammes.
- **Perspectives dans le cadre de FORUM**
 - Techniques hybrides (terminologiques et contextuelles)



Perspectives dans le cadre de FORUM

LIRMM

Perspectives

- Mettre en place des solutions **hybrides** prenant en compte :
 - les **aspects terminologiques** (fondées uniquement sur les termes),
 - les **aspects contextuels** (fondés sur la description des schémas, les labels des sous-arbres, *etc.*).
- Stage de Master Recherche de Yifei Tang.

Conclusions et Perspectives

Conclusions :

- **Prétraitements nécessaires.**
- **Techniques fondées sur les termes plus ou moins efficaces.**
- **Toutes les techniques terminologiques ont des limites souvent liées à la spécialité des lexiques utilisés et à l'absence d'informations syntaxiques.**

Perspectives :

- **Utilisation des informations contextuelles pour améliorer le résultat.**