CoRRecT : Démarche coopérative pour l'évaluation de systèmes de reconnaissance de termes

Chantal Enguehard

IRIN – Université de Nantes
2 rue de la Houssinière BP 92208 Nantes cedex 3, France
chantal.enguehard@irin.univ-nantes.fr

Résumé

La reconnaissance de termes dans les textes intervient dans de nombreux domaines du Traitement Automatique des Langues Naturelles, qu'il s'agisse d'indexation automatique, de traduction, ou d'extraction de connaissances. Nous présentons une méthodologie d'évaluation de Systèmes de Reconnaissance de Termes (SRT) qui vise à minimiser le temps d'expertise des spécialistes en faisant coopérer des SRT. La méthodologie est mise en œuvre sur des textes en anglais dans le domaine de la chimie des métaux et à l'aide de deux SRT : FASTR et SYRETE. Le banc de test construit selon cette méthodologie a permis de valider les SRT et d'évaluer leurs performances en termes de rappel et de précision.

Abstract

Recognizing terms in texts is useful in many Natural Language Processing applications : automatic indexation, summarization, translation, or knowledge extraction. We present a new methodology to evaluate Term Recognition Systems (TRS) so as to minimize the time required by experts to evaluate the results. This is done by making several TRS cooperate. This methodology is applied on English texts on metal chemistry with two systems : FASTR and SYRETE. The test bank we compiled evaluated the two systems and calculated the recall and precision rates.

Mots Clés :

Reconnaissance de termes, évaluation de la reconnaissance de termes, SRT, FASTR, SYRETE

Keywords

Recognition of terms, Term Recognition System, TRS, FASTR, SYRETE

1 Introduction

Seuls les spécialistes d'un domaine ont les connaissances langagières, techniques et pragmatiques nécessaires pour reconnaître des termes dans des textes, et donc pour effectuer une évaluation des systèmes. Mais pratiquement il est difficile de disposer de tels spécialistes, surtout pour des tâches longues (le corpus de test doit être grand), répétitives (il faut examiner les résultats de différents systèmes de reconnaissance de termes), et peu passionnantes. C'est pourquoi nous présentons une méthodologie de constitution des ressources nécessaires à l'évaluation de systèmes de reconnaissance de termes qui minimise le recours aux spécialistes. Nous détaillons la formalisation choisie pour les échanges de données. Nous montrons ensuite une application concrète de cette méthodologie à un corpus issu de la métallurgie.

2 Un banc de test pour la reconnaissance de termes

Un SRT (Système de Reconnaissance de Termes) reconnaît un ensemble de termes dans des textes composant un corpus et produit un corpus indexé. Dans ce modèle le corpus initial est « brut » dans le sens où il n'est pas nécessairement étiqueté.

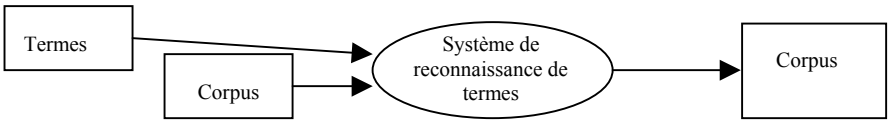

Figure 1 : Modélisation d'un Système de Reconnaissance de Termes

Une référence est un ensemble de termes et un corpus regroupant les textes dans lesquels les variantes de termes ont été préalablement reconnues. Ces deux ressources doivent être de taille importante afin de refléter au mieux la diversité des situations qui peuvent être rencontrées.

Nous définissons une approche itérative qui construit petit à petit le corpus de référence. A chaque pas de la construction du corpus

- un SRT fournit au banc de test un corpus indexé (les termes reconnus sont signalés) ;
- l'outil d'alignement compare ce corpus indexé par le SRT avec la référence dont il dispose. Il édite automatiquement des rapports destinés aux concepteurs des SRT : présentation des faux-positifs (termes reconnus de façon erronée), des faux-négatifs (termes non reconnus alors qu'ils devraient l'être) ainsi que des taux de rappel et de précision ;
- l'outil d'alignement produit des formulaires de validation où sont regroupées les variantes identifiées par le SRT qui ne figurent pas dans la référence ;
- les formulaires de validation sont examinés et remplis par les spécialistes du domaine ;
- l'outil d'intégration prend en compte le résultat des formulaires de validation et produit un nouveau corpus de référence dans lequel les variantes précédemment proposées à la validation sont signalées, ainsi que des formulaires de validation proposant les occurrences de variantes de termes ni validées ni invalidées par le spécialiste qui est intervenu.

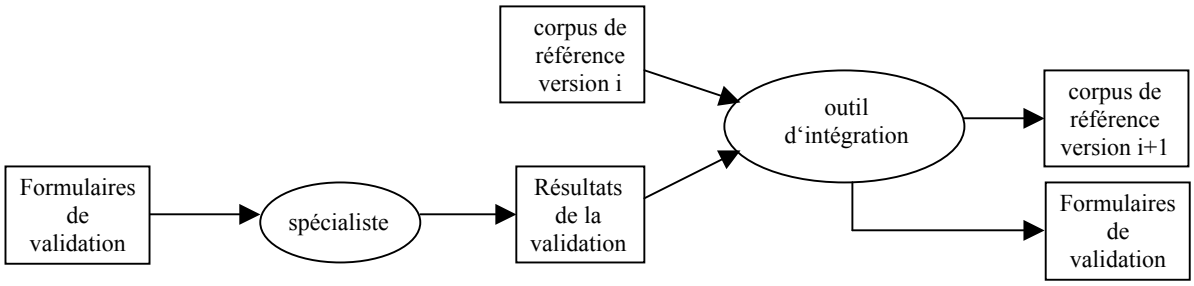

Figure 2 : Validation et outil d'intégration

L'application successive de cette chaîne de traitements avec plusieurs systèmes de reconnaissance de termes tend vers la définition d'un corpus de référence dans lequel le nombre de termes apparaissant dans le corpus de référence augmente. Le résultat ne dépend pas de l'ordre d'utilisation des différents SRT.

Les formulaires de validation proposés au spécialiste présentent chaque occurrence de variante de terme en contexte. Les experts ont pour consigne de valider l'occurrence si celle-ci « représente la même notion que le terme ». Le manque de précision de cette consigne reflète le défaut majeur de cette démarche : l'application finale de la reconnaissance de termes n'est pas explicitement définie, alors qu'elle peut influencer les performances attendues.

2.1 Formalisation

XML (http://www.xml.com) permet une définition précise des formats de données. Il facilite l'utilisation des ressources produites (le corpus de référence) par d'autres applications. Ce choix rend aussi le banc de test compatible avec les corpus existants et déjà formalisés à l'aide d'XML. Même si le corpus comprend d'autres informations signalées à l'intérieur de balises (par exemple des étiquettes morphosyntaxiques) le fonctionnement du banc de test ne sera pas perturbé.

2.1.1 Référence

La référence se divise en deux parties : la première regroupe les termes, la seconde le corpus de référence. Chacun des termes est défini par un identifiant (balise `<id>`), sa forme vedette (balise `<vedette>`) et, éventuellement, des informations supplémentaires (balise `<info>`). Le corpus de référence regroupe un ensemble de notices. Chaque identification de terme est signalée dans des balises `<variante>` : deux ancres signalent le début et la fin du terme ;. le statut de la variante est noté "Y"ou "N" selon la validation. La balise avis permet de recueillir d'éventuelles précisions de l'évaluateur.

```
<notice id="1">
  <variante refterme="5564" statut="Y" debut="ID12" fin="ID13">
    <avis>aucune variation</avis>
  </variante>
<texte>'Xi' is the dimensionless correlation length of the pair <ancre
ID="12"/>correlation function<ancre ID="13"/>. </texte></notice>
```

extrait des textes composant la référence

```
<terme id="5564">
  <vedette>Correlation function</vedette>
  <info>Fonction corrélation      N NH</info>
</terme>
```

extrait des termes composant la référence

2.1.2 Corpus indexé

Les SRT doivent fournir leurs résultats suivant un format fixé. Pour chaque notice identifiée par son numéro, apparaît la liste des identificateurs de termes reconnus et la notice elle-même. Chaque variante est identifiée par l'identifiant du terme reconnu et sa localisation signalée par des ancres

dans la notice. Chaque variante peut être accompagnée d'une information supplémentaire issue du SRT évalué.

3 Application de la méthodologie au corpus métal

3.1 Corpus

Le corpus est issu de l'Institut de l'Information Scientifique et Technique. Il est composé de 1280 résumés d'articles scientifiques dans le domaine de la chimie des métaux ainsi que d'une liste de 6582 termes du domaine (5239 termes complexes et 1343 termes simples). Les textes sont écrits en anglais et ont une étendue de 104 967 mots. Les termes sont assortis d'une traduction (approximative) en français, et d'un étiquetage morphosyntaxique, il n'y a pas de définition.

3.2 Deux systèmes de reconnaissance de termes

3.2.1 FASTR

La définition d'une typologie des variantes de termes complexes (composés de plusieurs mots) a permis de formaliser une grammaire qui exprime les transformations qui mènent d'un terme (sous sa forme normalisée) à ses variantes possibles. Les règles linguistiques ainsi définies sont implantées au sein du système FASTR (Jacquemin & Royauté 1995). FASTR utilise également des règles "négatives" permettant de conclure qu'une occurrence n'est pas une variante d'un terme.

3.2.2 SYRETE

La distance minimale d'édition entre deux chaînes de caractères est le nombre minimum d'insertions et de suppressions de symboles qui sont nécessaires pour transformer une chaîne en l'autre (Wagner 1974). Cette distance peut être pondérée par la somme des longueurs des deux chaînes. Cette mesure varie de 0, quand les chaînes sont strictement égales, à 1 quand les chaînes n'ont aucun symbole en commun. Elle peut être facilement adaptée à la reconnaissance de termes simples ou complexes (Enguehard 2000), elle est implantée au sein du système SYRETE (http://www.sciences.univ-nantes.fr/info/perso/permanents/enguehard/).

3.3 Résultats

FASTR identifie 4409 occurrences : selon ce système, 2934 sont des variantes de termes, 1475 n'en sont pas.

Comme le nombre d'occurrences à valider était très important, plusieurs personnes sont intervenues « en cascade » afin de minimiser le temps demandé au spécialiste du domaine. Bien que n'ayant aucune compétence particulière j'ai passé en revue toutes les occurrences et en ai validé (ou invalidé) plus de la moitié. Gloria Powell, qui est de langue maternelle anglaise, a ensuite passé en revue les 1743 occurrences restantes et en a laissé 603 qui ont finalement été proposées à Guy Saindrenan, spécialiste du domaine[1]. Cette procédure de validation a permis de construire la version 2 du corpus.

---

[1] Je remercie Guy Saindrenan et Gloria Powell sans qui ce projet n'aurait pu être mené.

SYRETE identifie 2818 occurrences de variantes de termes dont 1048 ne figurent pas dans la version 2 du corpus. La même procédure de validation des occurrences de termes a permis de produire la version 3 du corpus.

Finalement, l'outil d'alignement a été utilisé en prenant la version 3 du corpus comme référence.

Curieusement, le taux de rappel calculé pour les variantes de termes identifiées par FASTR est seulement de 63% alors que ce système est, a priori, très élaboré. Un examen approfondi des rapports édités a permis de détecter un dysfonctionnement : un grand nombre d'occurrences n'ont pas été identifiées car elles présentent des variations typographiques telles un changement dans la casse ou l'insertion d'un tiret. Le banc de test révèle ici son utilité comme outil de vérification du bon fonctionnement d'un SRT : bien que les résultats de FASTR aient été examinés à de nombreuses reprises, ce dysfonctionnement ne s'était jamais laissé soupçonner. Le taux de précision est de 89%. L'examen des résultats de FASTR révèle que deux transformations qui devraient, a priori, produire des variantes de termes entraînent en fait la reconnaissance d'occurrences qui ne sont pas des variantes de termes. Il s'agit de la transformation d'un adjectif en nom et de la transformation d'un nom en nom. Il apparaît que deux formes de variations sont particulièrement présentes : l'insertion d'un mot au sein d'une variante de terme est la forme de variation détectée par FASTR la plus courante (189 occurrences dont 77 % sont valides), et la permutation des mots composant une variante (111 occurrences dont 74 % sont valides). La prise en compte de ces phénomènes au sein d'un système de SRT traitant la langue anglaise améliore donc ses performances. SYRETE reconnaît les termes selon différentes modalités : la reconnaissance peut être sensible ou non à la casse, aux lettres accentuées et à la ponctuation. Elle peut être réalisée par l'opérateur d'égalité-souple, ou par l'égalité stricte de chaînes de caractères. Les expériences menées montrent que ces modalités influencent les taux de rappel tandis que les taux de précision gardent une valeur proche de 99 %. Les meilleurs résultats correspondent aux modalités les moins figées : dans ce cas SYRETE désaccentue le texte, le convertit en minuscule, élimine la ponctuation, et compare les chaînes de caractères avec l'opérateur d'égalité-souple. Comme le corpus étudié est de langue anglaise les accents jouent un rôle très faible.

4 Perspectives

La mise en œuvre de la méthodologie CoRRecT sur le corpus traitant de la chimie des métaux a permis de vérifier que la démarche suivie est réaliste mais que certaines difficultés subsistent. Tout d'abord, le processus intellectuel de validation n'est pas aisé, il est parfois difficile de déterminer si une occurrence est une variante d'un terme, il faudrait préciser les critères utilisés. Des expériences similaires d'évaluation par des spécialistes ont montré que leurs avis peuvent grandement diverger d'un spécialiste à l'autre (Daille et al. 1998). Nous projetons donc d'ajouter la possibilité de commenter les décisions prises et de les noter dans le corpus de référence.

La validation « en cascade » par plusieurs personnes de compétences croissantes, et les volumes traités (plusieurs milliers d'occurrences) laissent soupçonner des erreurs probables de validation. Nous projetons d'assortir les décisions notées dans le corpus d'un indice de confiance qui reflète les compétences de la personne qui a procédé à la validation, et de mettre en oeuvre un outil de visualisation du corpus de référence assorti d'un forum de discussion afin de réaliser des révisions des décisions de validation précédemment prises. L'évolution du format du corpus de référence par l'adjonction de nouvelles informations (comme un étiquetage morphosyntaxique des variantes, par exemple) pourra également être discutée dans ce forum.

Enfin, les systèmes évalués fournissent parfois des informations sur les occurrences de variantes de termes reconnues, par exemple FASTR assortit chaque reconnaissance d'une catégorie reflétant le

type de variation. Ces informations sont précieuses car elles peuvent éclairer les résultats. Ainsi, il serait intéressant d'observer les résultats de SYRETE afin de déterminer quels type de variations (au sens de FASTR) sont les mieux reconnues, ou les moins bien reconnues.

Nous avons donc étoffé la balise `<avis>` en lui ajoutant des attributs permettant de noter :

- l'identité et la qualification de la personne ayant validé la variante
- les informations fournies par les systèmes accompagnées de l'identité de ces systèmes

exemples :

```
<avis validation="Chantal Enguehard" qualification ="TALN">
<avis systeme="FASTR" information ="AtoV">
```

Une variante peut être accompagnée de plusieurs commentaires

5 Conclusion

Nous avons mis en œuvre la méthodologie CoRRecT de constitution d'un banc de test d'évaluation des systèmes de reconnaissance de termes qui minimise le temps de validation par des spécialistes du domaine. Nous avons constaté que les outils constituant le banc de tests fonctionnent correctement et avons évalué deux SRT. Cette expérience a montré que l'édition de rapports est essentielle pour détecter des dysfonctionnements éventuels de ces systèmes.

Par ailleurs, le corpus de référence constitue un ensemble de données de grande taille qui sont classées en deux groupes : les occurrences de variantes de termes, et celles qui n'en sont pas. Il peut constituer un ensemble d'entraînement pour des systèmes d'apprentissage fondés sur des exemples et des contrexemples.

Enfin, le corpus de référence ainsi que les outils afférents sont disponibles à l'adresse http://www.sciences.univ-nantes.fr/info/perso/permanents/enguehard/. Ce corpus, issu d'un processus de coopération entre différents systèmes de reconnaissance de termes, peut s'enrichir de nouvelles contributions que nous sollicitons vivement auprès des concepteurs de SRT.

5.1 Bibliographie

DAILLE B., GAUSSIER, É., LANGÉ, J.-M., 1998, "An Evaluation of Statistical Scores for Word Association", In *The Tbilissi Symposium on Logic Language and Computation: Selected Papers*, J. Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Levy, E. Vallduví, ed., p.177-188.

ENGUEHARD C., 2000, "Flexible-equality of terms: definition and evaluation", in Henrik L. Larsen, Janusz Kacprzyk, Slawonir Zadrozny, Troels Andreasen, et Henning Christiansen, réds., *Proceedings of the International Conference on Flexible Query Answering Systems*, p.289-300. ISBN 3-7908-1347-8.

JACQUEMIN C., ROYAUTÉ J., 1994, "Retrieving terms and their variants in a lexicalised unification-based framework", *Proceedings of the 17th annual international ACM SIGIR Conference on Research and Development in Information retrieval*, Dublin.

WAGNER R.A., FISCHER M.J., January, 1974, "The string-to-string correction problem". *J. of the Association for Computing Machinery*, 21 (1), p.168-173.

SYRETE : Variation des modalités

	Egalité-souple				Egalité stricte			
	Insensible à la casse		sensible à la casse		Insensible à la casse		sensible à la casse	
	! P	P	! P	P	! P	P	! P	P
Nombre de termes reconnus	2765	2349	2529	2269	2101	1713	1855	1619
rappel	68,21%	58%	62,44%	56,00%	52,13%	42,75%	46,02%	40,13%
précision	98,84%	98,93%	98,93%	98,9%	99,42%	99,36%	99,41%	99,32%

P : sensible à la ponctuation ! P : insensible à la ponctuation