
Spécifications de la reconnaissance des entités nommées et de leur utilisation

Romarc Besancon <romarc.besancon@cea.fr>

Copyright © 2005 Romarc Besançon - CEA-LIST

	Historique des versions	
Version 0.1	6 janv 2005	RB
	première rédaction	
Version 0.2	21 août 2006	GC
description de l'algorithme d'affectation de la catégorie et description de la configuration		

Table des matières

Introduction	1
Description	1
Reconnaissance des entités nommées	1
Utilisation des entités nommées	3
Enrichissement des entités nommées	4

Introduction

L'objectif de ce document est de spécifier le fonctionnement de la reconnaissance des entités nommées et de leur utilisation pour la recherche.

Description

Reconnaissance des entités nommées

On utilise le terme "entité nommée" pour désigner des noms propres, éventuellement composés, qui identifient des entités d'un type spécifique. Dans ce document, l'appellation est (abusivement) étendue pour prendre en compte d'autres types d'entités qui ne sont pas des noms propres, comme les nombres et les dates. Les types d'entités actuellement pris en compte par le système sont les suivants:

- TIMEX: dates, périodes
- NUMEX: nombres, valeurs numériques avec éventuellement leurs unités
- ORGANIZATION: noms d'organisations
- LOCATION: noms de lieux
- PERSON: noms de personnes
- PRODUCT: noms de produits
- EVENT: noms d'événements

Les entités nommées sont reconnues dans la phase de l'analyse linguistique. Cette reconnaissance a lieu après la désambiguïsation morphosyntaxique. Elle n'utilise pas les résultats de l'analyse syntaxique.

La reconnaissance s'effectue à l'aide d'un ensemble de règles spécifiques à chaque langue. Ces règles sont composées de règles contextuelles, utilisant des contextes particuliers pour repérer des entités nommées (par exemple, les titres ou les noms de professions comme annonceurs de noms de personnes), et de listes de noms connus (liste de prénoms, liste de noms de villes, de pays, de noms d'entreprises etc).

Après reconnaissance, une entité nommée est caractérisée par un type, une forme normalisée et l'ensemble des mots du texte qu'elle recouvre. Les formes normalisées associées aux entités nommées sont trouvées de la façon suivante:

- on peut spécifier dans les règles une forme normalisée particulière: dans ce cas, elle est gardée telle quelle;
 - on peut spécifier dans les règles l'appel à une fonction particulière pour construire dynamiquement une forme normalisée:
 - pour les noms de personnes, on applique une heuristique pour trouver le prénom et le nom;
 - pour les dates, on applique une heuristique pour identifier le jour, le mois et l'année. Actuellement, aucune inférence n'est faite pour trouver des valeurs sur des données relatives ou manquantes ("mardi dernier" ou "aujourd'hui" ne seront pas normalisés);
 - pour les nombres, on identifie l'unité si elle est spécifiée et on calcule la valeur numérique de l'expression (même si tout ou partie de l'expression est exprimée en lettres).
- Ces heuristiques peuvent dépendre des langues: elles ont été développées pour des langues latines qui ont un fonctionnement similaire (français, anglais, espagnol); elles n'ont pas été testées pour l'arabe ou le chinois.
- Si rien n'est spécifié, la forme normalisée de l'entité est la concaténation des formes normalisées des mots dont elle est composée.

Quand une entité nommée est détectée, l'ensemble des tokens qui la compose sont détachés du graphe d'analyse et remplacés par un nouveau token dont le lemme est la forme normalisée de l'entité et les propriétés linguistiques sont calculées d'après le type de l'entité, de manière configurable (Cf. section Configuration).

Performances de la reconnaissance

Ces mesures de performance sont données ici à titre indicatif: elles ont été établies dans le cadre du projet DETECT, sur les corpus étiquetés disponibles et avec la version de l'analyse linguistique livrée pour ce projet.

Tableau 1. Performance de la reconnaissance des entités nommées pour le Français, l'Anglais, l'Arabe et le Chinois (précision et rappel, par rapport à un étiquetage manuel de 5000 documents pour chaque langue)

	fre	eng	ara	chi
TOTAL	p=80.095% r=64.770%	p=59.264% r=46.350%	p=30.224% r=14.183%	p=39.227% r=3.709%
TIMEX	p=80.911% r=70.518%	p=82.490% r=73.684%	p=30.386% r=27.276%	p=54.217% r=5.478%
NUMEX	p=72.355% r=71.184%	p=45.881% r=74.133%	p=17.076% r=27.098%	p=88.235% r=1.061%
PERSON	p=85.243% r=63.827%	p=75.311% r=43.037%	p=88.717% r=9.034%	p=100.000% r=2.031%
LOCATION	p=72.355% r=71.184%	p=55.055% r=54.915%	p=38.500% r=34.535%	p=30.974% r=12.966%

Spécifications de la
reconnaissance des entités
nommées et de leur utilisation

	fre	eng	ara	chi
ORGANIZATION	p=83.809% r=45.520%	p=52.415% r=18.361%	p=0.000% r=0.000%	p=46.948% r=1.389%
PRODUCT	p=68.076% r=13.857%	p=65.823% r=0.840%	p=0.000% r=0.000%	p=0.000% r=0.000%
EVENT	p=47.073% r=26.395%	p=45.724% r=13.797%	p=0.000% r=0.000%	p=0.00 r=0.000%

Configuration

La reconnaissance des entités nommées se fait à plusieurs niveaux (voir aussi les documentations sur la configuration du système):

- S2-common.xml/common/CommonEntityTypes/SpecificEntities: liste des noms de types d'entités
- S2-common-<lng>.xml/LinguisticData/EntityTypes/TypeGroups, entrée SpecificEntities: Définit le groupe contenant la configuration des entités nommées pour cette langue (valeur habituelle: SpecificEntityTypes).
- S2-common-<lng>.xml/LinguisticData/EntityTypes/SpecificEntityTypes(habituellement, cf. ci-dessus)/TypeNames: liste des noms de types d'entités avec leur correspondance numérique
- S2-common-<lng>.xml/LinguisticData/EntityTypes/SpecificEntityTypes(habituellement, cf. ci-dessus)/Type2MicroMapping: micro-catégorie affectée au noeud remplaçant les noeuds couverts par l'entité si non spécifié dans la règle
- S2-common-<lng>.xml/LinguisticData/EntityTypes/SpecificEntityTypes(habituellement, cf. ci-dessus)/MappingWithCommonTypes: association entre les noms d'entités définis dans common et les noms locaux à la langue (habituellement égaux).

Utilisation des entités nommées

Les entités nommées reconnues sont utilisées dans l'indexation et la recherche, mais toute l'information n'est pas stockée.

Stockage des entités nommées dans l'index

Pour contourner certains problèmes de typage des entités (en particulier l'ambiguïté LOCATION/ORGANIZATION pour les noms de pays ou de villes lorsqu'ils sont utilisés pour représenter les organisations dirigeantes de ces pays/villes), le type des entités nommées n'est pas stocké dans l'index. Les entités nommées sont stockées dans l'index comme des mots composés, le fait qu'ils aient été produits à partir d'entités nommées n'est pas conservé.

Les catégories associées aux entités nommées (si l'index conserve les catégories) sont spécifiées dans le fichier de configuration de l'analyse linguistique (S2-lp-xxx.xml). Les catégories utilisées actuellement sont "nom commun" pour les nombres et dates, "nom propre" pour les autres.

Les parties des entités nommées sont également stockées dans l'index, avec la même catégorie.

Utilisation des entités nommées pour la recherche

Les entités nommées sont repérées dans les requêtes de la même façon que dans les documents. On cherche ensuite dans l'index, comme pour les autres mots, la présence de l'entité nommée (sans son type), ainsi que de ses parties.

Un poids spécifique peut être attribué aux entités nommées de la requête, en fonction de leur type. Dans ce cas, les entités nommées ont un poids relatif plus important par rapport aux autres mots

de la requête. Ces poids sont indiqués dans le fichier de configuration du moteur de recherche (S2-searchengine.xml).

Utilisation des entités nommées pour la visualisation des documents

Actuellement, les entités nommées n'étant pas stockées en tant que telles dans l'index, la visualisation de toutes les entités nommées présentes dans un document n'est pas possible.

Par contre, les entités nommées reconnues dans la requête sont conservées, et un surlignage particulier pour les mots du documents correspondant à une entité nommée trouvée dans la requête peut être envisagé.

Enrichissement des entités nommées

L'enrichissement doit se faire directement dans les fichiers sources des règles de reconnaissance des entités nommées.

L'ajout, la modification ou la suppression d'entités nommées prises en compte par des règles spécifiques fournies par l'utilisateur n'est pas possible actuellement.