Automatic Acquisition of Hyponyms from Large Text Corpora

Marti A. Hearst
University of California, Berkeley

Actes de COLING-92, Nantes, 23-28 Août 1992

BATTAGLIA Elena 30/01/2019

Hearst (1992). ***Automatic acquisition of hyponyms from large text corpora***

- méthode pour l'extraction automatique des **relations d'hyponymie** de **grands corpora textuels**

Domaine

- extraction des informations lexicales et sémantiques

Objectif général

- construction de lexiques hiérarchisés très larges, exploitables dans le TAL

Objectifs particuliers

- applicabilité de la méthode à tous types de texte
- éviter le besoin de prétraiter le texte, d'un outil de parsing très performant, d'une base de connaissance

Applications

extraire des relations d’hyponymie…pourquoi?

- enrichissement des lexiques, ex. WordNet
- interprétation sémantique de GN inconnus, ex. “broken bone” ~ “injury”
- recherche d’information
- reconnaissance de la similarité sémantique

Travaux antérieurs

- Alshawi (1987), Markowitz *et al.* (1986), Jensen & Binot (1987), Nakamura & Nagao (1988)
- extraction d'information lexicale des "**Machine readable dictionaries**" (MRDs)
- techniques de reconnaissance de relations *"pattern-based"* vs. *parsing*

✓ efficacité et précision linguistique

× limite intrinsèque — nombre des entrées du dictionnaire

↓

grands *corpora* textuels, une solution?

Technique de PATTERN RECOGNITION

idée de base

↓

associer la **relation lexicale** à des **constructions syntaxiques**

↓

PATTERNS

…sous trois conditions

- fréquents, récurrents dans de différents types de texte
- fiables, (presque) toujours pattern <> relation
- reconnaissables sans un traitement préalable du texte

Hyponymie entre lexique et syntaxe

"*The bowlute, <u>such as</u> the Bambara ndang, is plucked and has an individual curved neck for each string*"

- relation de hiérarchie entre deux lexèmes L_0 (**hyponyme**) et L_1 (**hyperonyme**)

L_0 is a kind of L_1

hyponym("Bambara Ndang", "bow lute")

- … instanciée par la construction syntaxique

NP_0, such as NP_1 (, NP_2…)

Démarche en quelques étapes

- identification des **patterns** qui manifestent la **relation d'hyponymie**
- description de la procédure pour les détecter automatiquement
- proposition d'integration des résultats dans un thesaurus "hand-built" (**WordNet**)
- exemple d'application de l'algorithme ⟶ validation de la méthode

ACQUISITION DES RELATIONS D'HYPONYMIE

Étape I. Identification des patrons lexico-syntaxiques

1. NP_0 such as {NP_1, NP_2 ... (and I or)} NP_N

The bow lute, such as the Bambara ndang

hyponym("Bambara dang", "bow lute")

2. such NP as {NP ,} * {(or | and)} NP

…works by such authors as Herrick, Goldsmith and Shakespeare

hyponym("Herrick", "author"), hyponym("Goldsmith", "author"), hyponym("Shakespeare", "author")

3. NP {, NP} * {,} or other NP

Bruises, wounds, broken bones and other injuries

hyponym("bruise", "injury"), hyponym("wound", "injury"), hyponym("broken bone", "injury")

…observation!

ACQUISITION DES RELATIONS D'HYPONYMIE

Étape I-II. Identification des patrons lexico-syntaxiques

4. NP {, NP}* {,} and other NP

…temples, treasuries, and other important civic buildings

hyponym("temple", "civic building"), hyponym("treasury", "civic building")

5. NP {,} including {NP, } * {or | and} NP

All common-law countries including Canada and England…

hyponym("Canada", "common-law country"), hyponym("Canada", "common-law country")

6. NP {,} especially {NP, } * {or | and} NP

…most european countries, especially France, England, and Spain.

hyponym("France", "european country"), hyponym("England", "european country"), hyponym("Spain", "european country")

→ comment identifier ces patrons?

ACQUISITION DES RELATIONS D'HYPONYMIE

Étape II. Détection automatique des patrons lexico-syntaxiques

- Choisir une relation d'hypoymie. Ex: membre/groupe
- Collecter une liste des termes (à partir d'un lexique ou des patrons observés). Ex: "England-country", "tank-vehicle"
- Enregistrer les environnements syntaxiques où les deux termes de la relation co-occurrent
- Trouver les points communs entre ces environnements ⟶ formaliser un patron lexico-syntaxique
- Utiliser le nouveau patron pour repérer d'autres exemples de la relation dans le texte et repartir du deuxième point.

ACQUISITION DES RELATIONS D’HYPONYMIE

Étape III. Intégration des résultats dans WordNet

Que-ce que WordNet?

"... is an attempt to organize lexical information in terms of word meanings, rather than word forms. In that respect, WordNet resembles a thesaurus more than a dictionary ..."

- exemple de thésaurus exploité pour des tâches de TAL
- réseau de **synsets** (groupes de synonymes), organisé de façon hiérarchique, sur la base des relations d’hyponymie
- contient surtout des noms non modifiés
- dans quelle mesure les relations “hyponym(N_0, N_1)” repérées se retrouvent dans WordNet?

→ comparaison > intégration > amélioration

ACQUISITION DES RELATIONS D'HYPONYMIE

Étape III. Intégration des résultats dans WordNet

Trois situations possibles

- VERIFY

  ✓ N_0 et N_1 ✓ hyponym(N_0, N_1)

- CRITIQUE

  ✓ N_0 et N_1 × hyponym(N_0, N_1)

- AUGMENT

  × N_0 et N_1 × hyponym(N_0, N_1)

intégrer les noms et les relations manquants n'est pas toujours évident!

ACQUISITION DES RELATIONS D'HYPONYMIE

Résultats et évaluation

Application de l'algorithme d'acquisition

- au pattern 1: NP_0 such as {NP_1, NP_2 ... (and I or)} NP_N
- restrictions sur les membres de la relation: noms non modifiés (= WordNet)
- texte: Grolier's American Academic Encyclopaedia (8.6M mots)

Vérification des relations dans WordNet

- 152 relations trouvés > Wordnet…..mais performance à améliorer!
- bonne qualité
- difficultés récurrentes: métonymie, sous-spécification, dépendance du contexte, hyperonyme très général, rôle des modificateurs

ACQUISITION DES RELATIONS D'HYPONYMIE

Résultats et évaluation

cereals:	rice* wheat*
countries:	Cuba Vietnam France*
hydrocarbon:	ethylene
substances:	bromine* hydrogen*
protozoa:	paramecium
liqueurs:	anisette* absinthe*
rocks:	granite*
substances:	phosphorus* nitrogen*
species:	steatornis oilbirds
bivalves:	scallop*
fungi:	smuts* rusts*
fabrics:	acrylics* nylon* silk*
antibiotics:	ampicillin erythromycin*
institutions:	temples king
seabirds:	penguins albatross*
flatworms:	tapeworms planaria

flatworms:	tapeworms planaria
amphibians:	frogs*
waterfowl:	ducks
legumes:	lentils* beans* nuts
organisms:	horsetails ferns mosses
rivers:	Sevier Carson Humboldt
fruit:	olives* grapes*
hydrocarbons:	benzene gasoline
ideologies:	liberalism conservatism
industries:	steel iron shoes
minerals:	pyrite* galena
phenomena:	lightning*
infection:	meningitis
dyes:	quercitron

Bilan

Avantages

- détection automatique des patrons <> petit nombre de termes au départ
- **vs.** méthodes d'apprentissage sur base statistique <> grand nombre de relations exprimées
- peu ou pas d'analyse préalable du texte <> méthode peu onéreuse

Questions "ouvertes"

- pertinence des modificateurs > quelles restrictions imposer?
- polysémie des termes de la relation > comment désambiguïser?
- insertion de nouveaux termes > dans quel synset? (pour WordNet)

…nécessité de tests ultérieurs!

MERCI!