

Treball de recerca Intel·ligència Artificial

Anotació automàtica de recursos Web mitjançant patrons lingüístics

El procés d'annotació consisteix en afegir informació semàntica a contingut textual. Principalment, es tracta d'associar una categoria conceptual d'un determinat domini (contingudes en una ontologia) a cadascuna de les entitats amb nom (instàncies) d'un text. Per exemple, donat un recurs textual que parli sobre components d'ordinador i l'ontologia del domini associada, es tractaria d'etiquetar entitats com “*Intel Core 2 Duo*” o “*AMD Athlon X2*” com a *processadors x86* o “*ATI Radeon 9800 Pro*” o “*Nvidia Geforce 7800 GT*” com a *targes gràfiques 3D*.

L'annotació Web és un procés molt important del qual en depèn l'èxit de la Web Semàntica i permetria el desenvolupament d'eines d'Extracció i Recuperació d'Informació molt més avançades que els buscadors actuals. No obstant, la quantitat de pàgines web disponibles fa que no sigui factible una anotació manual dels recursos. És per això que s'estan desenvolupant diferents aproximacions per tal d'automatitzar – encara que sigui parcialment – el procés d'annotació.

Des d'un punt de vista general, existeixen tres aproximacions bàsiques a l'annotació automàtica segons la seva generalitat:

- La més clàssica consisteix en aprendre a identificar i classificar un conjunt reduït de categories força regulars i ben delimitades com són *noms de persones, llocs, dates, noms d'organitzacions, etc.* Típicament es basa en algorismes d'aprenentatge supervisats que utilitzen grans quantitats d'exemples prèviament anotats. Tot i proporcionar bons resultats, el seu àmbit d'aplicació es força limitat degut al reduït nombre de categories considerades (ni tant sols es podria considerar una meta-ontologia).
- En un nivell superior trobem aproximacions que intenten aprendre a identificar cadascun dels conceptes continguts a una ontologia a base de definir a aprendre regles o funcions que les identifiquin. Existeixen aproximacions supervisades – basades en exemples prèviament anotats – i no supervisades – basades en expressions regulars associades a cada concepte -. Les primeres tenen el problema de la poca escalabilitat, degut a la gran quantitat de possibles categories a aprendre i dels exemples anotats manualment que cal proporcionar. Les segones presenten la dificultat de la definició manual i estricta d'expressions regulars per a cada concepte. Això no sempre es factible doncs resulta molt complicat trobar regularitats estructurals en quant a l'expressió textual.
- Finalment trobem les aproximacions que no depenen d'un conjunt predefinit o limitat de categories i que intenten ser capaces d'identificar i etiquetar qualsevol entitat de qualsevol concepte de qualsevol ontologia. Aquests sistemes es basen en un conjunt de regles o patrons dependents del llenguatge que típicament expressen especialitzacions d'un concepte. Per exemple, donat el patró per a l'anglès “*NP such as NP or NP*”, podem trobar l'expressió “*... graphic cards such as ATI Radeon 9800 Pro or Nvidia Geforce 7800 GT*”. Si el concepte “*graphic card*” és un concepte de l'ontologia per exemple de “*components hardware d'ordinadors*”, podem anotar les entitats trobades amb l'etiqueta corresponent. Degut a la manca de supervisió d'aquestes aproximacions i de la possible ambigüitat del llenguatge, molt sovint fan servir algun tipus d'anàlisi estadístics per tal d'inferir si l'etiqueta conceptual amb que s'etiquetarà una entitat és realment adient.

En aquest treball de recerca es tracta de desenvolupar una metodologia d'anotació automàtica de recursos Web d'aquest últim tipus. Des d'un punt de vista molt general, caldria realitzar els següents passos:

- Es parteix d'una ontologia de qualsevol domini i d'un conjunt de planes web (que es podrien obtenir a partir de l'ontologia i crides a un cercador web).
- Analitzar el contingut de cada plana per tal d'identificar entitats amb nom. Es poden fer servir un conjunt de patrons lingüístics per tal d'identificar a quina categoria poden pertànyer. També cal identificar adequadament el nom complet (ex. “ATI Radeon 9800 Pro”) de cada entitat.
- Realitzar un anàlisi estadístic per a cadascun dels candidats a entitat etiquetada fent servir la Web com a base de coneixement. Concretament, es poden formular *queries* a cercadors web per tal d'obtenir la rellevància d'una relació (ATI Radeon 9800 Pro és una *targeta gràfica*) a partir del recompte de *hits* retornant pel cercador. D'aquesta forma es podria acabar de decidir si és adient etiquetar l'entitat amb la categoria corresponent. L'obtenció d'estadístics a partir de crides a cercadors web és molt interessant doncs en permet obtenir d'una forma molt immediata mesures de correlació basades en la distribució de la informació del repositori més gran disponible.
- Un cop finalitzat el procés per tots els candidats de la plana, anotar-la fent servir algun llenguatge estàndard (ex: RDF(s)).

Les principals tasques d'aquest TR són:

- Estudi de l'estat de l'art en anotació automàtica de recursos Web. Es proporcionaran les principals referències actualitzades.
- Disseny de la metodologia d'anotació basada en patrons del llenguatge i estadístics obtinguts de la Web. Més concretament caldrà considerar:
 - o Mecanismes per identificació de entitats amb nom dins d'un text. Caldrà ser capaç de detectar el nom complet de cada candidat. Això també es pot fer mitjançant *queries* a cercadors que ens informin de la forma més habitual en que es presenta un determinat candidat (ex: majúscules, acompanyat per determinants, etc.).
 - o Mecanismes per associar el candidat al concepte de l'ontologia més adequat. Es podran fer servir patrons dependents del llenguatge que expressin aquestes relacions.
 - o Estudi de les possibilitats que ofereixen els principals cercadors web disponibles en quant a llenguatge de *queries* (ex: possibilitat de distingir majúscules i minúscules, operadors tipus NEAR, NOT, etc.). En funció d'aquest estudi, dissenyar les *queries* que ens permetran portar a terme les metodologies desenvolupades per identificar entitats amb nom i associar-les a les categories de l'ontologia segons la seva rellevància.
- Implementació de la metodologia, utilitzant algun llenguatge d'anotació estàndard. Caldrà estudiar els entorns d'anotació que ja existeixen per tal de facilitar el treball. També caldrà fer servir diverses eines de processament bàsic de llenguatge natural, *pàrsing* dels resultats dels cercadors web (Google, MSNSearch, Altavista), etc. El tutor proporcionarà les pautes, eines i exemples pertinents.
- Estudi detallat i avaluació dels resultats obtinguts, per tal de demostrar la viabilitat de la solució, la influència dels paràmetres utilitats, variacions en les aproximacions dissenyades, etc.