Lingüística computacional

Antoni Oliver González (coordinador)
Xavier Gómez Guinovart
Irene Castellón Masalles

PID_00146603

Universitat Oberta de Catalunya
www.uoc.edu

Antoni Oliver González

Professor dels Estudis d'Arts i Humanitats de la Universitat Oberta de Catalunya i director acadèmic del postgrau de Traducció i tecnologies. És doctor en Lingüística, llicenciat en Filologia Eslava i enginyer tècnic de Telecomunicacions. La seva àrea de recerca se centra en l'aplicació de tècniques de processament del llenguatge natural a les tasques de traducció.

Xavier Gómez Guinovart

Professor de Lingüística computacional a la Universitat de Vigo, i responsable del Seminari de Lingüística Informàtica i del Grup de Tecnologies i Aplicacions de la Llengua Gallega d'aquesta universitat. La seva recerca se centra en el processament del llenguatge natural i en les aplicacions lingüístiques de la informàtica.

Irene Castellón Masalles

Professora del Departament de Lingüística General de la Universitat de Barcelona i coordinadora general del programa de doctorat Ciència Cognitiva i Llenguatge. És doctora i llicenciada en Filologia Romànica. La seva àrea de recerca se centra en la lingüística computacional, especialment en els nivells sintàctic i semàntic.

Primera edició: febrer 2010

Av. Tibidabo, 39-43, 08035 Barcelona
Disseny: Manel Andreu
Realització editorial: Eureca Media, SL
Dipòsit legal: B-8.976-2010

Introducció

Aquests materials, corresponents a l'assignatura *Lingüística computacional*, s'han preparat especialment per a oferir una introducció a aquesta matèria amb un enfocament pràctic. Totes les tècniques de processament del llenguatge natural que es presenten s'acompanyen de programes que les implementen. D'aquesta manera es poden veure les diferents tècniques en funcionament, cosa que permet percebre de manera molt més clara els resultats que s'assoleixen, les dades lingüístiques necessàries i les aplicacions pràctiques que tenen. Això no vol dir que s'hagin deixat de banda els aspectes teòrics bàsics de la disciplina. Aquests aspectes també es tracten amb profunditat i la seva presentació es veu reforçada pels programes que s'ofereixen.

En el mòdul 1 s'ofereix una introducció general a la lingüística computacional. Es presenten les diferents línies d'investigació i les principals aplicacions. La lingüística computacional és una àrea molt àmplia i és impossible presentar totes les tècniques i els formalismes associats en aquest manual. Per aquest motiu la lectura d'aquest mòdul serà interessant per a tenir una visió general de la disciplina.

Per a poder executar i modificar els programes que presentem en aquests materials no és necessari tenir un bon nivell previ de programació. Tot i això, en el mòdul 2 s'ofereix una introducció al llenguatge de programació Python. La lectura i la realització dels exercicis proposats en aquest mòdul no ha d'espantar ningú. Tot i que pugui semblar molt complicat i que no s'assoleixin tots els objectius d'aquest mòdul, serà possible seguir amb profit la resta d'exemples pràctics que es presenten en aquests materials. Qui assoleixi un cert nivell de programació en aquest mòdul podrà aprofitar els seus coneixements no solament en aquesta assignatura, si no també en el seu treball diari amb l'ordinador, ja que serà capaç de fer petits programes que poden automatitzar tasques habituals i repetitives. S'ha escollit el llenguatge Python per diversos motius: és un llenguatge modern, potent i fàcil d'aprendre. A més, és multiplataforma i tot el necessari per a programar en el Pyhton és gratuït i de lliure distribució. El fet que sigui multiplataforma vol dir que els programes que fem es podran executar sense problemes tant en el Windows com en el Linux i el Mac (sempre que no facin servir mòduls dependents d'un determinat sistema operatiu). Una altra característica interessant és que es poden trobar nombrosos paquets, mòduls i programes lliures en el Python que ens facilitin el desenvolupament de les nostres aplicacions. Un d'aquests paquets disponibles i que farem servir de manera intensiva en aquest curs serà l'NLTK (Natural Language Toolkit). Aquest paquet proporciona un seguit de mòduls i classes que implementen d'una manera eficient molts dels processos necessaris en processament del llenguatge natural.

El mòdul 3 tracta del processament de corpus textuals. En aquest mòdul aprendrem a obrir fitxers de text, atenent-ne la codificació, i fer càlculs i processaments bàsics. Entre els càlculs que farem hi ha el càlcul de freqüències absolutes i relatives, *n*-grames, col·locacions, etc. Aprendrem també a tractar corpus anotats i analitzats. Veurem què és el WordNet i com podem utilitzar-lo des dels nostres programes en el Python.

A partir d'aquest moment i en els mòduls posteriors aprendrem tècniques de processament del llenguatge natural en diferents nivells. Així, en el mòdul 4 aprendrem a tractar la morfologia de les paraules i veurem els diferents formalismes que s'han desenvolupat en morfologia computacional. En el mòdul 5 veurem les diferents tècniques per a portar a terme la tasca anomenada *etiquetatge morfosintàctic*. Aquesta tasca consisteix a associar a cada paraula d'un text una etiqueta morfosintàctica i de manera opcional el lema associat. En el mòdul 6 aprendrem a fer anàlisis fragmentals, és a dir, determinar els constituents sintàctics d'una oració sense establir les relacions entre aquests constituents.

El mòdul 7 és un parèntesi en què aprendrem els fonaments d'un altre llenguatge de programació: el Prolog. Aquest llenguatge ha estat molt emprat en intel·ligència artificial, ja que planteja un paradigma diferent de programació, en què no es programen les solucions a un problema, sinó que s'especifica el problema d'una manera lògica i es deixa la solució a l'ordinador.

En el mòdul 8 presentem els principals formalismes sintàctics i de manera especial les gramàtiques lliures de context. Aprendre a escriure gramàtiques en aquest formalisme i veurem els diferents tipus d'analitzadors que ens permeten obtenir les anàlisis de les oracions a partir d'aquestes gramàtiques. També veurem les gramàtiques lliures de context probabilístiques, que són unes gramàtiques que tenen una informació addicional: la probabilitat de cada una de les produccions de la gramàtica. Finalment, en aquest mòdul, veurem les gramàtiques de dependències, que ofereixen un tipus d'anàlisi diferent i se centren a establir com les paraules es relacionen les unes amb les altres.

Per acabar, en el mòdul 9, presentem una sèrie d'eines per al processament del català, que estan disponibles amb una llicència lliure. Aquestes eines, a més de proporcionar-nos la possibilitat de fer diferents tasques, ens proporcionen també una sèrie de recursos lingüístics que podrem fer servir en les nostres pròpies aplicacions.

Objectius

Els objectius generals d'aquesta assignatura són els següents:

1. Tenir una panoràmica general de la disciplina.

2. Conèixer les principals tècniques i processos de la lingüística computacional.

3. Conèixer els diferents nivells d'anàlisi lingüística que es poden portar a terme amb tècniques de lingüística computacional.

4. Conèixer el paquet NLTK (Natural Language Toolkit).

5. Saber executar i modificar petits programes en el Python relacionats amb el processament del llenguatge natural.

6. Ser capaços d'instal·lar i executar diferents aplicacions relacionades amb el processament del llenguatge natural.

Continguts

Mòdul didàctic 4

Morfologia computacional

Antoni Oliver

1. Fonaments lingüístics
2. Definició i objectius de la morfologia computacional
3. Tècniques i formalismes en morfologia computacional
4. Aprenentatge de la morfologia
5. Els programes Automorphology i Linguistica

Mòdul didàctic 5

Etiquetatge morfosintàctic

Antoni Oliver

1. Els etiquetadors morfosintàctics
2. Tècniques per a l'etiquetatge morfosintàctic
3. Un analitzador morfològic basat en un diccionari
4. Desambiguació mitjançant regles creades manualment
5. Desambiguació mitjançant tècniques estadístiques
6. L'etiquetador de Brill
7. Tècniques d'aprenentatge automàtic aplicades a l'etiquetatge morfosintàctic
8. Avaluació pràctica d'etiquetadors
9. Alguns etiquetadors disponibles

Mòdul didàctic 6

Anàlisi fragmental (*chunking*)

Antoni Oliver

1. El concepte de *chunk*, anàlisi fragmental i anàlisi sintàctica superficial
2. Un *chunker* senzill
3. Chinking
4. Representació de *chunks*: etiquetes i arbres
5. Creació i avaluació d'un *chunker* a partir de corpus
6. Aplicacions dels analitzadors fragmentals i dels analitzadors superficials

Mòdul didàctic 7

Introducció a la programació en Prolog

Irene Castellón

Mòdul didàctic 8

Formalismes sintàctics

Antoni Oliver

1. Gramàtiques lliures de context
2. Analitzadors per a gramàtiques lliures de context
3. Gramàtica de trets
4. Les gramàtiques lliures de context probabilístiques
5. Gramàtiques de dependències

Mòdul didàctic 9

Eines i recursos per al català i castellà

Antoni Oliver

1. El Corrector
2. DACCO
3. Freeling
4. Apertium

Annex

Introducció al Natural Language Toolkit (NLTK)

Antoni Oliver

Bibliografia

Cada mòdul s'acompanya d'una bibliografia específica. Tot i això, hi ha una sèrie d'obres de referència que són de consulta recomanada. En podem destacar les següents:

Bird, S.; Klein E.; Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly.

Nugues, M. P. (2006). *An Introduction to Language Processing with Perl and Prolog*. Springer.

Mitkov, R. (ed.) (2003). *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Hauser, R. (2001). *Foundations of Computational Linguistics*. Springer.

Manning, C. D.; Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.