TERMINOLOGIAREN ERAUZKETA AUTOMATIKOA ETA BERE APLIKAZIOA EUSKARARAKO

Izaskun Aldezabal, Iñaki Alegria, Xabier Artola, Nerea Ezeiza, Ruben Urizar

Informatika Fakultatea, Euskal Herriko Unibertsitatea
P.K. 649. 20080 DONOSTIA
acpalloi@si.ehu.es

1 Sarrera

Azken urteotan testu teknikoetatik terminologia automatikoki eraucteko tresnak ari dira garatzen zenbait hizkuntzatarako, ingeleserako batez ere. Lengoaia Naturalaren Prozesamenduan (NLP) eta beste eremutan egindako ikerketen ondorioz teknologia prest dago horrelako aplikazioak garatu ahal izateko, nahiz eta emaitzak ikusita oraindik giza-laguntza behar izan, automatikoki hautatutako terminologiaren artean azken aukeraketa egiteko.

Tresna horien aplikazio-eremuak askotarikoak dira baina bi multzo handitan bana daitezke: informazioa indexatzeko (text indexing) eta terminoen glosategiak eraikitzeko. Lehen eremua puri-purian dago, Internet dela-eta informazio asko dago eskuragarri baina informazio horren selekzioan dago erabilgarritasunaren gakoa. Terminoen glosategiak automatikoki eraiki ahal izatea oso mesedegarria da terminologiaren, itzulpenaren zein argitarapenen munduan. Gainera, terminologia oso modu dinamikoan bilakatzen den arloetan, informatikan adib., halako tresnarik gabe ia ezinezkoa da lan terminologiko eraginkorrik egitea.

Halako tresna bat euskararako garatu nahi badugu, eragozpen gehiago topatuko dugu ondoko hiru arrazoiengatik: batetik, bateratze-prozesua bukatzeke dagoenez, terminologia finkatzea konplexuagoa da; bestetik, egindako ikerketak murritzak dira; eta, azkenik, hizkuntza eranskaria izateagatik, terminologia identifikatzeko behar den tratamendua inguruko hizkuntzena baino konplexuagoa izango da.

2 Terminologiaren erauzketa

Bibliografia aztertuz nabari daiteke gai honi buruz azken urteotan egon den interes-hazkundea. Hazkunde horren adibide gisa, tresna hauek aipa daitezke: *LEXTER* (Bourigault, 92), AT&Tko *Termight* (Church & Dagan, 94), IBMko *TERMS* (Justeson & Katz, 95) eta *NPtool* (Arpper, 95).

Terminoaren definizio formal eta osoa lortzea lan neketsua da eta horretan datza lanen atal garrantzitsu bat: terminoen ezaugarriak mugatzea. Corpusetatik termino teknikoak lortzeko konbinatu ohi dira NLPko teknikak (ezagumendu linguistikoan oinarritutakoak) eta teknika estatistikoak.

Ingelesez egile gehienek izen-sintagmetara murrizten dituzte termino teknikoak.

2.1 Teknika linguistikoak

Teknika linguistikoak erabiltzen dira batez ere terminoen hasierako selekzioa egiteko. Horretarako, eredu morfosintaktikoak erabili ohi direnez gero, komenigarria da testua analizaturik edukitzea edo gutxienez etiketatua. Analisi sintaktiko sakona eskatzen ez bada ere, azaleko analisi sintaktiko sendoa funtsezkoa da hautapena patroi sintaktikoen arabera egingo bada. Tresna linguistikoen kalitateak baldintzatuko ditu, hein handi batean behintzat, tresnaren emaitzak. Hala ere, proiektu batzuetan ez da analisi morfologikorik edo sintaktikorik egiten (Su *et al.*, 96) eta hitz-bikoteak edota hirukoteak aukera daitezke bestelako murrizpenik gabe. Beste sistema batzuetan tarteko irtenbideak hartzen dira, adib. lista batean agertzen diren hitzak bereizgarritzat hartzen dira.

Hasierako selekzioaz gain, ezagumendu linguistikoa funtsezkoa da terminoen normalizazioan ere; termino batzuk beste luzeago batzuen baitan egon daitezkeenez, haien artean diskriminatu egin behar da, eta horretarako informazio morfosintaktikoa inportantea izan badaiteke ere, maiz formula estatistikoak soilik erabili ohi dira.

Analisi morfologikoarekin eta desanbiguazioarekin lotuta dago lematizazioa. Inguruko hizkuntza batzuetan (ingelesa, frantsesa eta espainera adib.) hitz- forma hutsak, edo gehienez numero bereizturik, edukitzea nahikoa den bitartean flexio konplexuko hizkuntzetan horrek emaitza kaxkarrak ekarriko lituzke, eta lematizazioa ezinbesteko aurreprozesua izango da. Adibidez, euskaraz informatikan erabiltzen den *sistema eragile* terminoa identifikatzeko *sistema eragilea*, *sistema eragilearen*, *sistema eragileko*, *sistema eragileei*, *sistema eragilera* eta beste forma asko bildu beharko dira. Hala ere, azaleko sintaxia ezinbestekoa izango da lematizazio hutsa ez baita nahikoa. Adib., eta aurreko adibideari helduz, *sistemaren eragile* ez da *sistema eragile* terminoaren flexioa.

2.2 Teknika estatistikoak

Eredu linguistikoari jarraitzen dioten balizko terminoak murrizteko erabilohi dira metodo estatistikoak proiektu gehienetan.

Aplikatutako metodoak asko aldatzen dira proiektuaren arabera: sinpleena izango litzateke maiztasun absolutu minimo bat eskatzea (Justeson & Katz, 95), baina orokorrean formula probabilistiko anitz konbinatzen dira eta formula horien artean gailentzen dena *mutual information* deitutakoa dugu (Church & Hanks, 90). Metodo horren bidez bi osagaien arteko korrelazioa neurtzen da ondoko formulaz:

$$MI(a,b) = \log_2 (P(a,b) / P(a)\ P(b))$$

non P(a) eta P(b) dira osagai bakoitzeko agerpen-probabilitatea corpusean, eta P(a,b) da jarraian edo gertu agertzeko duten probabilitatea.

Formula hau oinarritzat hartzen da sistema gehienetan, baina batzuetan konbinatzen da sofistikatuagoekin. Horrela proiektu batean (Su *et al.*, 96) testu partikular batean agertzen diren probabilitateak corpus handi eta orekatu baten probabilitateekin alderatzen dira maiztasun erlatiboak lortzeko, hauek funtsezkoak izan baitaitezke terminologia diskrimatzeko orduan.

Aipatutako teminoen normalizazioa (beste luzeagoetan agertzen direnak murrizteko) Manchesterreko Unibertsitateko proiektu batean (Frantzi & Ananiadou, 96) *C-value* izeneko formula berri bat proposatzen dute. Formula hori aplikatuz gai dira bereizten *soft contact lenses, hard contact lenses,* eta *contact lenses* termino gisa eta, aldiz, baztertzen *soft contact*.

Hala eta guztiz ere, formula hauekin aritzen direnak ohartzen dira estatistikaren mugaz eta informazio semantikoaren beharraz jabetzen dira. Adibidez, Ferrari-k eta Prince-k (Ferrari & Prince, 96) dioten bezala, *mutual information* izeneko formulan ezin dira kontutan hartu sinonimia bezalako fenomenoak eta horren ondorioz emaitzak ez dira behar diren bezain doiak. Semantikak oso paper garrantzitsua joka dezake terminologiaren erauzketan bai erabakiak hartzeko orduan bai irteera sare semantiko gisa antolatzeko.

2.3 Emaitzak

Lortzen diren emaitzak ez dira oraindik beharko liratekeenak erauzketa zeharo automatikoa egiteko. Oreka bilatu behar da estaldura (teknikoki *recall* esaten zaio) eta doitasunaren artean (*precision*). Lortzen diren terminoak benetako terminoak izan daitezen, kalitate onekoak alegia, estaldura jaitsi egingo da eta termino asko zerrendatik at geratuko dira. Eta alderantziz, estaldura bultzatzen baldin badugu, termino tekniko ez diren espresio asko azalduko dira, doitasunaren kaltetan.

EZAUGARRIA	LEXTER (Bourigault, 92)	TERMS (Justenson, 95)	(Frantzi & Ananiadou, 96)	(Su *et al.*, 96)
HAUTAKETA	sintaktikoa (oso azalekoa)	sintaktikoa	sintaktikoa (adj|ize)*ize	bigramak edo trigramak
ETIKETATZEA	BAI	BAI	BAI (Brill)	BAI
LEMATIZAZIOA	EZ	EZ	EZ	BAI
OINARRIZKO ESTATISTIKA	maiztasuna	maiztasuna	maiztasuna	mutual information
BESTE ESTATISTIKAK	EZ	EZ	C-value	maiztasun erlatiboa
SEMANTIKA	EZ	EZ	EZ	EZ

ESTALDURA/ DOITASUNA	%95 / ??	?? / %85	%93/%45 %30/%82	%96/%48 %97/%40

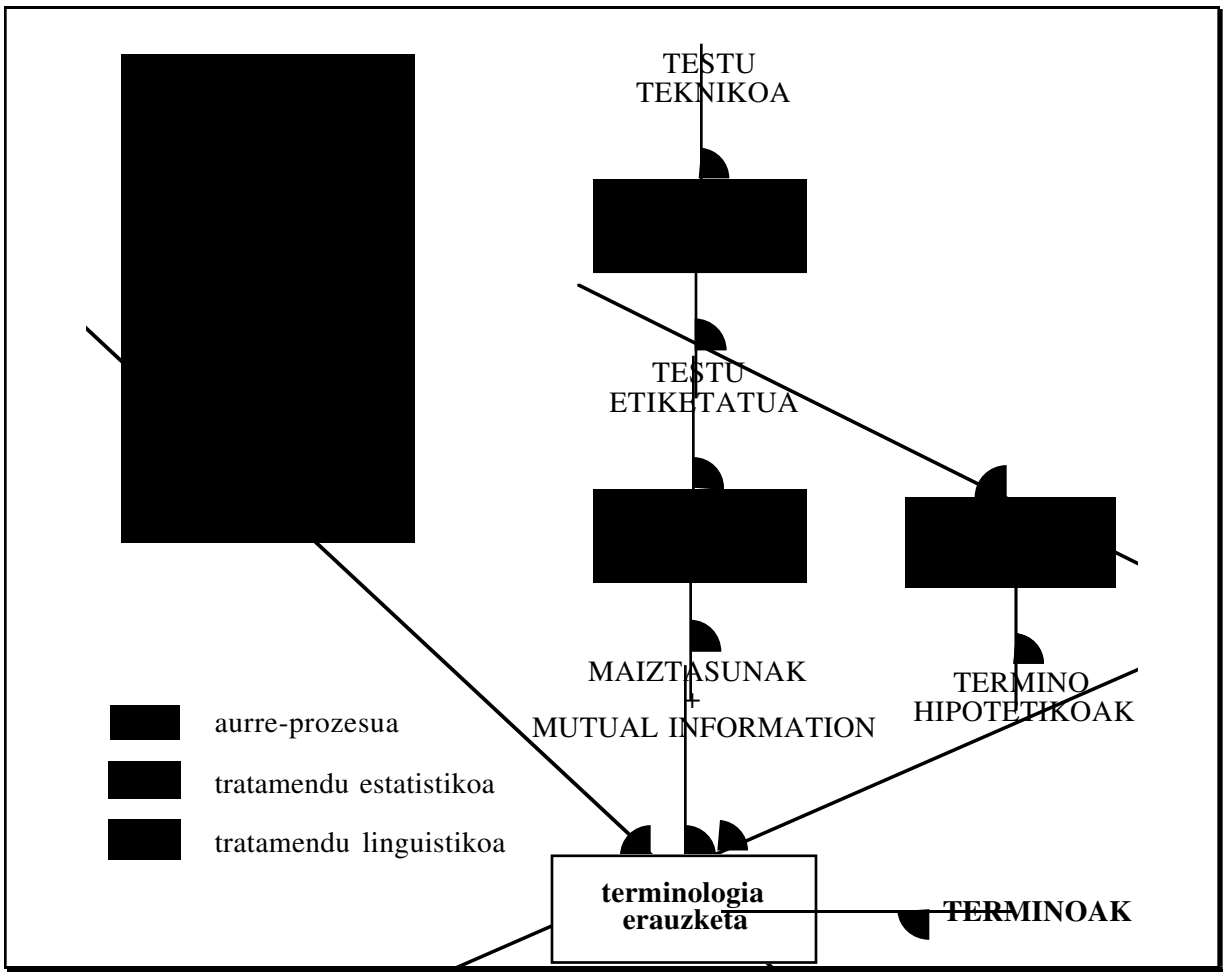
1.irudia.- Zenbait sistemaren ezaugarriak

Erdibide horretan estaldurari lehentasuna ematen zaio atzetik terminologia murrizteko pertsona bat baldin badago. %95 inguruko estaldura lortzeko doitasuna %50era jaitsi ohi da, eta doitasuna %85 ingurukoa izan dadin estaldura %35era ere ez da iristen.

1. irudian azaltzen da aztertutako sistemen ezaugarrien laburpena. Bertan azaltzen da *artearen egoera* gaur egun, nahiz erabilitako tekniken inguruan nahiz lortutako emaitzen aldetik.

3 Euskararako aplikazioa

IXA taldearen asmoa da mota honetako tresna bat garatzea euskararako. Horretarako analizatzaile morfologikoa jadanik prest dago (Alegria *et al.*, 96), lematizatzaile/etiketatzaile bat bukatzear dago (Aduriz *et al.*, 96) eta azaleko sintaxiari ere ekin diogu. Azpimarratzekoa da tresna hauek duten ezaugarrietako bat, sendotasuna hain zuzen, posiblea baita analizatzea, etiketatzea eta lematizatzea edozein hitz, bere lema lexikoan egon ez arren. Honen bidez lor daiteke edozein testutako hitz guztiak analizatzea, etiketatzea eta lematizatzea lexikoa eguneratzeko beharra izan gabe.



2.irudia.- Proposatutako sistemaren arkitektura

Tresna horiek prest dauden bitartean termino teknikoen modelizazioari ekin behar diogu, hau da, murriztu behar ditugu termino teknikoen ezaugarriak. Horretarako, dauden hiztegi teknikoetan oinarritu, eta teknika estatistikoak erabiliz, eredu nagusiak lortu behar dira. Emaitzarik ez badugu ere, eredua izen-sintagmarena baino zabalagoa izango dela susmatzen dugu. Beste aldetik, izen-sintagmetan termino teknikoak hautatzerakoan barneko deklinabide-kasua erabakigarria izan daitekeelakoan gaude, hau da, aurreko *sistema eragile*aren adibidean azaldu den bezala, beti ez da lema kontutan hartu beharko batzuetan osagai baten forma osoa izango baita terminoaren parte.

Tresna honetarako aurreikusten dugun arkitektura 2. irudian azaltzen dena da, eta oraingoz ez dugu ezagumendu semantikoaren erabilpena aurreikusi. Bertan azaltzen denaren arabera, corpus orokor handi batetik abiatu behar gara, eta corpusean agertzen diren hitzak, lemak eta hauei dagozkien kategoriak kontuan hartuz erreferentziazko estatistikak kalkulatuko dira. Estatistika hauek oinarria izango dira testu teknikoetan agertzen diren maiztasun ez-ohizkoak detektatu ahal izateko. Teknika hau ez da erabiltzen, bibliografian aztertutako kasuetan behinik behin, baina uste dugu lagungarria izan daitekeela doitasun ona lortzeko.

Aurreprozesu hori eginda edukiz gero, testu batetik (liburu, artikulu, eskuliburu, etab.) terminologia erauzi nahi dugunean testu “tekniko” hori analizatu behar da, lematizazioa, etiketatzea eta azaleko sintaxia lortzeko, eta behin bere osagaiak bereiztu direnean metodo estatistikoak aplikatuko dira, testu barneko korrelazioak kalkulatzeko (*mutual information*) zein hizkuntza orokorrekiko dituen desbiderapenak lortzeko. Emaitza hauetan oinarrituta tratatutako testuaren terminologia osatzen duten elementuak lortuko dira.

Erreferentziak:

(Alegria *et al.*, 96) Alegria I., Artola X., Sarasola K., Urkia M. Automatic Morphological Analysis of Basque. *Literary and Linguistic Computing* **11** (4): 193-203. Oxford University Press. 1996.

(Aduriz *et al.*, 96) Aduriz I., Aldezabal I., Alegria I., Artola X., Ezeiza N., Urizar R. EUSLEM: A lemmatiser/tagger for Basque. *Proc. of the Euralex´96*, Goteborg, Sweden. 1996.

(Ananiadou, 94) Ananiadou S. A Methodology for Automatic Term Recognition. *Proc. of the Conference on Computational Linguistics (Coling-94)*, 1034-1038, Kyoto, Japan. 1994.

(Arpper, 95) Arpper A. Term Extraction from Unrestricted Text. *http://www.lingsoft.fi/doc/nptool/term-extraction.html*.

(Bourigault, 92) Bourigault D. Surface grammatical analysis for the extraction of terminological noun phrases. *Proc. of the Conference on Computational Linguistics (Coling-92)*, 977-981, Nantes, France. 1992.

(Church & Hanks, 90) Church K., Hanks P. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics* **16**:22-29. 1990.

(Church & Dagan, 94) Church K., Dagan I. Termight: Identifying and Traslating Technical Terminology. *Proc. of the 4th Conference on Applied Language Processing*, Stuttgart, Germany. 1994.

(Daille *et al.*, 94) Daille B., Gaussier E., Lange J. Towards Automatic Extraction of Monolingual and Bilingual Terminology. *Proc. of the Conference on Computational Linguistics (Coling-94)*, 515-521, Kyoto, Japan. 1994.

(Ferrari & Prince, 96) Ferrari K.T., PrinceS. Création et Extension Automatiques de Dictionnaries Terminologiques Multilingues Specialisés a partir de Corpus Monolingues. *Proc. of the Conference on Natural Language Processing and Industrial Applications,* 79-86. Moncton, N.B. Canada. 1996.

(Frantzi & Ananiadou, 96) Frantzi K.T., Ananiadou S. Extracting Nested Collocations. *Proc. of the Conference on Computational Linguistics (Coling-96)*, 41-46. 1996.

(Justeson & Katz, 95) Justeson J.S., Katz S.M. Technical Terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* **1** (1): 9-27. Cambridge University Press. 1995.

(Oueslati *et al.*, 96) Oueslati R., Frath P., Rousselot F. Term Identification and Knowledge Extraction. *Proc. of the Natural Language Processing and Industrial Applications (NLP+IA)*, 191-196, Moncton, Canada. 1996.

(Su *et al.*, 96) Su K., Wu M., Chang J. A Corpus-based Aproach to Automatic Compound Extraction. *Proc. of the Conference on Computational Linguistics (Coling-96)*, 243-247. 1996.