TERMINOLOGIAREN ERAUZKETA AUTOMATIKOA ETA BERE APLIKAZIOA EUSKARARAKO

Izaskun ALDEZABAL
Iñaki ALEGRIA
Xabier ARTOLA
Nerea EZEIZA
Ruben URIZAR

Informatika Fakultatea. EHU

1. Sarrera

Azken urteotan testu teknikoetatik terminologia automatikoki erauzteko 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 numeroa 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 erabili ohi 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:

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

non P(a) eta P(b) diren osagai bakoitzeko agerpen-probabilitatea corpusean, eta P(a,b) den 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 diskriminatzeko orduan.

Aipatutako terminoen normalizazioa (beste luzeago batzuetan agertzen direnak murrizteko) Manchesterreko Unibertsitateko proiektu batean (Frantzi & Ananiadou, 96) *C-value* izeneko formula berri bat proposatzen dute. Formula hori aplikatuz gai dira bereizteko *soft contact lenses*, *hard contact lenses*, eta *contact lenses* termino gisa eta, aldiz, baztertzeko *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 kontuan hartu sinonimia bezalako fenomenoak eta horren ondorioz emaitzak ez dira behar 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 gauzak zertan diren 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, posible baita edozein hitz analizatzea, etiketatzea eta lematizatzea, 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 alde batetik, 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 bakarrik kontuan 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 abiatuko 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 orokorrarekiko 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.

****

LABURPENA / RESUMEN / RÉSUMÉ / ABSTRACT

Terminologiaren erauzketa automatikoa eta bere aplikazioa euskararako

1. Sarrera

Azken urteotan testu teknikoetatik terminologia automatikoki erauzteko tresnak ari dira garatzen zenbait hizkuntzatarako, baina oraindik giza laguntza behar izaten da automatikoki hautatutako terminologiaren artean azken aukeraketa egiteko. Horren adibide gisa, tresna hauek aipa daitezke: *LEXTER* (Bourigault, 92) AT&Tko *Termight* (Church & Dagan, 94) IBMko *TERMS* (Justeson & Katz, 95) *NPtool* (Arpper, 95).

Aplikazio-eremuak bi multzo handitan bana daitezke: informazioa indexatzeko eta terminoen glosategiak eraikitzeko. Gainera, terminologia oso modu dinamikoan bilakatzen den arloetan, informatikan adib., halako tresnarik gabe ia ezinezkoa da lan terminologiko eraginkorrak egitea.

Halako tresna bat euskararako garatu nahi badugu, eragozpen gehiago topatuko dugu ondoko hiru arrazoiengatik: bateratze-prozesua bukatzeke izateagatik, egindako ikerketak murritzak direlako eta hizkuntza eranskaria izateagatik.

2. Terminologiaren erauzketa

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.

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. 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).

Analisi morfologikoarekin eta desanbiguazioarekin lotuta dago lematizazioa. Flexio konplexuko hizkuntzetan hitz-forma bakarrik tratatzeak emaitza kaxkarrak ekarriko ditu eta lematizazioa ezinbestekoa izango da. Ezagumendu linguistikoa funtsezkoa da terminoen normalizazioan ere; termino batzuk beste luzeago batzuen baitan egon daitezkeenez, haien artean diskriminatu egin behar baita.

2.2 .Teknika estatistikoak

Eredu linguistikoari jarraitzen dioten balizko terminoak murrizteko erabili ohi 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.

2.3. Emaitzak

Lortzen diren emaitzak ez dira oraindik beharko liratekeenak erauzketa zeharo automatikoa egiteko. Oreka bilatu behar da estaldura (*recall*) eta doitasunaren artean (*precision*). Oreka horretan estaldurari lehentasuna ematen zaio atzetik terminologia murrizteko pertsona bat badago. % 95 inguruko estaldura lortzeko doitasuna % 50-era jaitsi ohi da, eta doitasuna % 85 ingurukoa izan dadin estaldura % 35era ere ez da iristen.

3. Euskararako aplikazioa

IXA taldearen asmoa da euskararako mota honetako tresna bat garatzea. 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.

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, termino teknikoak hautatzerakoan barneko deklinabide-kasua erabakigarria izan daiteke.

El vaciado terminológico automático y su aplicación para el euskera

1. Introducción

En los últimos años se están desarrollando en varias lenguas instrumentos para efectuar vaciados terminológicos automáticos de textos técnicos, si bien todavía se hace necesaria la intervención humana para hacer la última selección de los términos elegidos automáticamente. Como ejemplo de lo anterior pueden citarse los siguientes instrumentos: *LEXTER* (Bourigault, 92), *AT & Tko Terminght* (Church & Dagan, 94) *TERMS* de IBM (Justeson & Katz, 95) *NPtool (*Arpper, 95).

Pueden dividirse en dos grandes grupos las áreas de aplicación: área de indexación de la información y área de confección de glosarios terminológicos. Además, en las áreas en las que la terminología evoluciona de modo dinámico, como por ejemplo la informática, sin ese tipo de instrumental resulta prácticamente imposible llevar a cabo un trabajo terminológico efectivo.

Si pretendemos desarrollar un instrumento similar para el euskera, toparemos con mayores inconvenientes debido a estas razones: el proceso unificador de la lengua no ha concluido, las investigaciones efectuadas son limitadas y, por último, el euskara es una lengua aglutinante.

2. Vaciado terminológico

Es una ardua labor conseguir una definición formal y completa de un término y en eso consiste precisamente un apartado importante del trabajo: definir las características de los términos. Para conseguir del corpus términos técnicos se suelen combinar las técnicas NLP (basadas en el conocimiento lingüístico) y las técnicas estadísticas.

2.1. Técnicas lingüísticas

Las técnicas lingüísticas se emplean fundamentalmente para efectuar la selección inicial de los términos. Como se suelen emplear modelos morfosintácticos, resulta conveniente tener analizado el texto o, por lo menos, etiquetado. La calidad de la herramienta lingüística condicionará en gran medida por lo menos los resultados de la misma. De todos modos, en algunos proyectos no se efectúa ni análisis morfológico ni sintáctico. (Su et al., 96).

La lematización está ligada al análisis morfológico y a la desambiguación. En las lenguas de flexión compleja, el tratar solamente el aspecto formal de las palabras acarreará malos resultados y será necesaria la lematización. El conocimiento lingüístico también es primordial en la normalización terminológica; ya que como algunos términos pueden formar parte de otras unidades más largas, se ha de efectuar una discriminación entre ellos.

2.2. Técnicas estadísticas

En la mayoría de los proyectos, los métodos estadísticos se han venido empleando para reducir los supuestos términos que siguen el modelo lingüístico. Los métodos aplicados varían mucho en función del proyecto, por lo que lo más simple sería pedir una frecuencia absoluta mínima (Justeson & Katz, 95), si bien, en general, se combinan numerosas fórmulas probabilísticas.

2.3. Resultados

Los resultados que se obtienen no son aún los que se precisarían para efectuar un vaciado absolutamente automático. Se ha de encontrar el equilibrio entre la cobertura

(recall) y la precisión *(precision)*. En ese equilibrio se le otorga preferencia a la cobertura, siempre que haya una persona que lleve a cabo la reducción terminológica. Para obtener una cobertura del 95% se suele reducir la precisión al 50%, y para que la precisión ronde el 85%, la cobertura no se reduce ni al 35% siquiera.

3. Aplicación al euskera

El grupo IXA tiene la intención de desarrollar una herramienta de este tipo para el euskera. Para ello, ya está preparado el analizador morfológico (Alegria et al., 96), el lematizador/etiquetador está a punto de finalizarse (Aduriz et al., 96) y también estamos trabajando la sintaxis del nivel superficial.

Mientras se preparan dichas herramientas, habremos de trabajar sobre la modelización de los términos técnicos, es decir, hemos de reducir las características de los mismos. Con tal fin, basándonos en los diccionarios técnicos existentes y empleando técnicas estadísticas, se han de conseguir modelos principales. Aunque aún no contamos con resultados, intuimos que el modelo será más amplio que el del sintagma nominal. Por otra parte, en la elección de términos técnicos, el caso de declinación interna puede resultar decisivo.

Le dépouillement terminologique automatique et son application pour l'euskera

1. Introduction

Lors des dernières années des instruments sont en train de se développer dans plusieurs langues afin de réaliser les dépouillement terminologiques automatiques des textes techniques, si bien aujourd'hui encore l'intervention de l'homme est nécessaire pour faire la dernière sélection des termes choisis automatiquement. Comme exemple de ce qui précède nous pouvons citer les instruments suivants: *LEXTER (Bourigault, 92), AT & Tko Terminght (Church & Dagan, 94), TERMS de IBM (Justeson & Katz, 95), NPtool (Arpper, 95).*

Les domaines d'application peuvent être divisés en deux grands groupes: le domaine d'indexation de l'information et le domaine de confection de glossaires terminologiques. De plus, dans les domaines dans lesquels la terminologie évolue de façon dynamique, comme par exemple l'informatique, sans ce type d'instrument il est pratiquement impossible de mener à bien un travail terminologique effectif.

Si nous prétendons développer un instrument similaire pour l'euskera, nous nous trouverons face à des inconvénients étant donné ces deux raisons: le processus unificateur de la langue n'est pas terminé, les recherches effectuées sont limitées et, finalement, l'euskera est une langue agglutinante.

2. Dépouillement terminologique

Obtenir une définition formelle et complète d'un terme est un travail ardu, et c'est précisément en cela que consiste une section importante du programme: définir les caractéristiques des termes. Pour obtenir des termes techniques du corpus linguistique on combine généralement les techniques NLP (basées sur la connaissance linguistique) et les techniques statistiques.

2.1. Techniques linguistiques

Les techniques linguistiques sont fondamentalement employées pour réaliser la sélection initiale des termes. Comme on emploie généralement des modèles morphosyntaxiques, il convient que le texte soit analysé ou, du moins, qu'il soit étiqueté. La

qualité de l'outil linguistique conditionnera en grande mesure au moins les résultats de celle-ci. De toute façon, dans certains projets on n'effectue ni l'analyse morphologique ni l'analyse syntaxique. (Su et al., 96).

La lemmatisation est liée à l'analyse morphologique et à la désambiguïsation. Dans les langues de flexion complexe, le fait de traiter simplement l'aspect formel des mots entraînera des mauvais résultats et la lemmatisation sera nécessaire. La connaissance linguistique est également primordiale dans la normalisation terminologique; comme de nombreux termes peuvent faire partie d'autres unités plus longues, il faut effectuer une discrimination entre eux.

2.2. Techniques statistiques

Dans la plupart des projets, les méthodes statistiques ont été employées pour réduire les termes hypothétiques qui suivent le modèle linguistique. Les méthodes appliquées varient beaucoup en fonction du projet, et donc le plus simple serait de demander une fréquence absolue minimum (Justesson & Katz, 95), même si, en général, de nombreuses formules probabilistes sont employées.

2.3. Résultats

Les résultats obtenus ne sont pas encore ceux dont on aurait besoin pour réaliser un dépouillement absolument automatique. Il faut trouver l'équilibre entre la couverture *(recall)* et la précision *(precision)*. Dans cet équilibre, la couverture a une préférence, pourvu qu'il y ait une personne qui mène à bien la réduction terminologique. Pour obtenir une couverture de 95% on réduit généralement la précision à 50%, et pour que la précision soit autour de 85% la couverture n'est même pas réduite à 35%.

3. Application à l'euskera

Le groupe IXA a l'intention de développer un outil de ce type pour l'euskera. Pour cela, l'analyseur morphologique est déjà prêt (Alegria et al., 96), le lemmatisateur/étiquetteur est sur le point d'être terminé (Aduriz et al., 96) et nous avons également travaillé la syntaxe au niveau superficiel.

Pendant que nous préparons ces outils, nous devrons travailler sur la modélisation des termes techniques, c'est à dire, nous devons réduire les caractéristiques de ces derniers. Dans ce but, en nous basant sur les dictionnaires techniques existants et en employant des techniques statistiques, nous devons obtenir des modèles principaux. Bien que nous n'ayons pas encore de résultats, nous devinons que le modèle sera plus large que celui du syntagme nominal. D'autre part, dans le choix de termes techniques, le cas de la déclinaison interne peut être décisif.

Automatic terminology extraction and its application to Basque

1. Introduction

In recent years work has begun to develop instruments in several languages for automatic terminology extraction in technical texts, though human intervention is still required to make the final selection from the terms automatically chosen. As an example we can cite the following instruments: *LEXTER* (Bourigault, 92), *AT & Tko Terminght* (Church & Dagan, 94), *TERMS* by IBM (Justeson & Katz, 95) and *NPtool* (Arpper, 95).

Their areas of application can be divided into two main groups: information indexing and the making-up of terminological glossaries. In areas where terminology is developing

dynamically, such as computer science, it is almost impossible to carry out effective terminological work without an instrument of this type.

If a similar instrument is to be developed for Basque we shall come up against more major drawbacks, because the unifying process of the language has not been completed, research carried out is limited and Basque is an agglutinative language.

2. Terminology extraction

It is a hard task to obtain a formal, complete definition of a term, but that is precisely what a major part of this work consists of: defining the characteristics of terms. To obtain technical terms from the corpus a combination of NLP techniques (based on linguistic knowledge) and statistical techniques is usually used.

2.1. Linguistic Techniques

Linguistic techniques are used basically to make the initial selection of terms. Morpho-syntactic models are usually used, so it is advisable to have the text already analysed or at least labelled. The results are conditioned heavily by the quality of the linguistic tool used. In any event in some projects neither morphological nor syntactic analysis is carried out (Su et al., 96).

Lemmatisation is linked to morphological analysis and the removal of ambiguities. In complex inflected languages poor results will ensue if only the formal aspect of words is dealt with: lemmatisation will be necessary. Linguistic knowledge is also of prime importance in the standardisation of terminology: a discrimination between terms must be made, because some of them may form part of longer units.

2.2. Statistical Techniques

In most projects statistical methods have been used to reduce the assumed terms which follow the linguistic model. The methods applied vary widely from project to project, so the simplest idea is to require a minimum absolute frequency (Justeson & Katz, 95), though several probabilistic formulae are generally combined.

2.3. Results

The results obtained are not yet those required for absolutely automatic extraction. A balance must be found between recall and precision. In this balance preference is given to recall, provided there is a person who can carry out the terminology reduction. To obtain a recall of 95% precision is usually reduced to 50%, and for a precision of 85% cover is not reduced even to 35%.

3. Application to Basque

The IXA Group intends to develop a tool of this type for Basque. The morphological analyser is already being prepared (Alegria et al, 96), the lemmatizer/labeller is almost completed (Aduriz et al, 96) and work has been done on surface level syntax.

While these tools are being prepared, we must work on the modelling of technical terms, i.e. we must reduce their characteristics. To that end, basing work on existing technical dictionaries and using statistical techniques, principal models must be obtained. We do not yet have any results, but we believe that the model will be wider than the noun phrase. In the choice of technical terms, the case of internal declension may prove decisive.