

Elements Textuals per a Processament del Llenguatge Natural

1. ANTECEDENTS DE LA RECERCA QUE ES PRETÉN REALITZAR	2
2. OBJECTIUS	3
3. METODOLOGIA	4
4. PLA DE TREBALL I CALENDARI D'EXECUCIÓ	6
BIBLIOGRAFIA	7

1. Antecedents de la recerca que es pretén realitzar

En els darrers anys, els mètodes basats en dades (data-driven methods) han introduït un gir paradigmàtic en PLN. Mitjançant tècniques d'inducció a partir de textos associats a anàlisis lingüístics (corpus anotats), s'han obtingut models de llenguatge probabilístics que han introduït millores quantitatives i qualitatives en aplicacions com l'etiquetatge morfològic o l'anàlisi sintàctic.

No obstant, aquestes tècniques gairebé no s'han aplicat a nivell de discurs, degut a l'escassetat de textos anotats amb informació discursiva de suficient longitud. En els darrers anys s'estan desenvolupant cada cop més corpus anotats a nivell discursiu (Carlson et al. 2003, The Potsdam Commentary Corpus), i hi ha diversos treballs en què s'han induït analitzadors a partir d'aquests corpus (Marcu i Echihi 2002, Reitter 2003). No obstant, aquests treballs es basen totalment en una teoria de l'organització dels textos, la Rhetorical Structure Theory (Mann i Thompson 1988), que ha rebut multitud de crítiques des dels seus inicis per la seva manca d'adequació als fenòmens que hom troba als textos i a les necessitats de representació del discurs pròpies del PLN.

El fet que no hi hagi corpus anotats amb una teoria del discurs àmpliament acceptada per la comunitat científica es deu justament a la manca d'una teoria d'aquestes característiques. En comparació amb el consens en la representació dels fenòmens lingüístics en fonologia, morfologia o sintaxi, a nivell discursiu manca encara establir pràcticament totes les qüestions fonamentals sobre l'organització dels textos: quines són les unitats bàsiques, què les identifica, com es relacionen entre elles, quins mecanismes lingüístics serveixen per crear estructura discursiva i com ho fan, etc. Malgrat la multitud de propostes per respondre aquestes preguntes, cap no aconsegueix donar una explicació àmpliament satisfactòria als fenòmens discursius més bàsics.

Davant d'aquest desgavell teòric, els mètodes basats en dades suposen una sòlida via per donar compte dels fenòmens lingüístics. En la línia de les corrents estructuralistes més estrictes, l'anàlisi formal del discurs reconeix la utilitat de les dades explícites, més que no pas les interpretacions subjectives, com a indicadors de com s'organitza l'espai semàntic de les relacions discursives. Més concretament, en lingüística computacional, Knott (1996) proposa explotar els connectors discursius de l'anglès (paraules com *however*, *because*) com a marques explícites de relacions discursives entre segments textuais. Així, estableix un mètode bàsic per obtenir una organització jeràrquica dels connectors discursius basada en les seves relacions de sinonímia i hiperonímia. La interpretació d'aquesta organització permet a l'autor d'identificar primitives de significat discursiu que actuen com a trets distintius dins d'aquesta jerarquia.

La metodologia bàsica proposada per Knott és parcialment automatitzada per Hutchinson (2005), que desenvolupa algorismes per determinar relacions entre connectors discursius mitjançant la comparació dels seus contextos d'ocurrència, amb una precisió considerablement alta.

A la meua tesi doctoral (Alonso Alemany 2005) vaig desenvolupar un conjunt de primitives de significat discursiu amb una aproximació que està en la mateixa línia que les anteriors, però amb les següents diferències:

- A diferència de Knott, em baso exclusivament en evidència textual explícita i reconeixible mitjançant tècniques de PLN superficials.
- A diferència de Knott i Hutchinson, treballa paral·lelament en anglès, castellà i català, mitjançant un conjunt de connectors discursius (o marcadors discursius) parcialment equivalents en les tres llengües.
- Aplicant tècniques de clustering, descobreixo característiques redundants en una descripció preliminar dels marcadors discursius (Alonso et al. 2002).
- Aplicant tècniques d'adquisició de lèxic, determino la utilitat relativa de diverses característiques textuais superficials per la caracterització de marcadors del discurs (Alonso et al. 2002).
- Aplicant la teoria de mapes semàntics amb un petit lèxic trilingüe de marcadors discursius, descobreixo àrees de significat cohesionades per l'anglès, castellà i català.
- Aplico les primitives de significat per l'anàlisi manual d'un petit corpus, aconseguint un acord acceptable entre jutges (Alonso et al. 2004).
- Mostro com aquestes primitives de significat són útils per explicar un model de rellevància textual compartit per diversos jutges humans en un experiment de resum de textos.
- Mostro com aquestes primitives de significat són útils en aplicacions de resum automàtic de textos. (Alonso i Fuentes 2002, Alonso i Fuentes 2003, Alonso et al. 2004).

Així doncs, el meu treball aplica els principis bàsics de Knott i Hutchinson, però amb diferències orientades a generar recursos per al processament automàtic a nivell discursiu de les llengües catalana i castellana. L'estat actual de la recerca en aquestes dues llengües fa que les diferències entre ambdós mètodes siguin substancials. Per una banda, l'estudi de l'organització del discurs, i sobretot dels marcadors discursius, es encara més incomplet que per l'anglès, per la qual cosa una funció important de l'estudi basat en corpus dels marcadors és proporcionar dades per millorar aquesta descripció. Per altra banda, l'anàlisi automàtica de les llengües catalana i castellana és encara molt superficial, de manera que qualsevol mètode de tractament de la llengua s'hi ha d'adequar.

2. Objectius

La meua recerca té un objectiu teòric i un altre aplicat. Des d'un punt de vista teòric, pretenc obtenir un conjunt de relacions discursives bàsic vàlid interlingüísticament, a partir de l'evidència que es troba en els textos a nivell superficial. A nivell aplicat, pretenc crear recursos d'àmplia cobertura per l'anàlisi automàtic dels textos mitjançant aquestes relacions i l'explotació de l'evidència textual en què es basen.

Aquests objectius generals es desglossen en els següents sub-objectius:

1. Caracteritzar el comportament d'aquells elements textuais (ETs) que expliciten l'organització discursiva dels textos, per les llengües anglesa, francesa, catalana i castellana, començant per un conjunt de marcadors del discurs prototípic (amb paraules como *perquè*, *però*, *a més*) i augmentant aquest conjunt mitjançant tècniques d'exploració contextual. Aquesta caracterització es realitzarà de manera automàtica amb supervisió manual.
2. Per cada una de les quatre llengües treballades, establir classes d'equivalència i de relacions de substituïbilitat entre ETs, per crear una taxonomia a partir de la qual obtindrem la taxonomia de les relacions discursives expressada per cada una de les classes d'equivalència.
3. Unir les quatre taxonomies monolingües, creant una única taxonomia vàlida per les quatre llengües, mitjançant judicis de substituïbilitat basats en el mètode de mapes semàntics. Es prendrà el conjunt de relacions d'aquesta taxonomia més general como el conjunt estàndard de relacions per analitzar text.
4. Desenvolupar un analitzador discursiu que pugui donar una certa anàlisi dels textos explotant la ocurrència d'ETs com els caracteritzats a 1., mitjançant la seva semàntica associada en la taxonomia obtinguda a 3.
5. Millorar l'anàlisi dels textos obtingut automàticament amb una anàlisi humana que es basi en el conjunt de relacions discursives obtingut a 3. El resultat d'aquesta millora serà un petit corpus anotat manualment. D'aquest corpus s'obtindran:
 - a) nous ETs, que seran incorporats a la taxonomia (3.), i a l'analitzador (4.)
 - b) noves gramàtiques, que s'incorporaran a l'analitzador discursiu (4.)Aquest pas serà repetit iterativament, amb l'objectiu d'obtenir un analitzador discursiu prou fiable i un corpus anotat de bona mida.
6. Un objectiu important del projecte és la difusió dels resultats obtinguts i les metodologies desenvolupades, per tal de contribuir a establir un marc de consens sobre els fenòmens discursius, sobretot en llengües diferents de l'anglès. D'altra banda, els recursos generats a partir d'aquest projecte seran de lliure distribució (licència GPL).

3. Metodologia

En aquesta secció especifiquem els mètodes a seguir pel desenvolupament dels sub-objectius llistats a la secció anterior.

1. Per caracteritzar el comportament dels ETs discursius, ja sigui del conjunt inicial o dels que s'obtinguin mitjançant el sub-objectiu 5.b), obtindrem un gran nombre d'ocurrències de cada ET en text real, juntament amb el seu context. Aquestes ocurrències s'obtidran o bé de corpus de la llengua ja establerts o bé de la world wide web.
 - 1.1. cada ocurrència d'un ET i el seu context seran analitzats superficialment mitjançant l'analitzador Freeling, per identificar el lema i la categoria morfològica de les paraules que co-ocorren amb l'ET en qüestió, així com els grups sintagmàtics que formen.
 - 1.2. si ja disposem d'un analitzador discursiu, cada ocurrència serà també analitzada discursivament.
 - 1.3. determinarem quines característiques del context són rellevants per la caracterització d'un ET donat. Per fer-ho aplicarem tècniques d'exploració contextual en la línia de Wonsever i Minel (2004). També seran considerades rellevants aquelles característiques amb major informació mútua amb l'ET en qüestió. Per poder obtenir la informació mútua entre cada ET i el seu context, realitzarem un estudi preliminar en què determinarem:
 - a) quines característiques del context són susceptibles de ser rellevants per la caracterització d'ETs,
 - b) com es distribueixen aquestes característiques en el text, per poder contrastar la seva distribució en text amb la seva distribució en presència d'un ET determinat.Aquest estudi pot aportar informació molt valuosa per completar i aprofundir la descripció gramatical dels ETs
 - 1.4. per cada una de les característiques del context susceptibles de ser rellevants per la caracterització d'ETs, determinarem manualment si aporta informació discursiva per si mateixa, i per tant podem considerar-la com un ET. Així, ampliarem el conjunt inicial d'ETs, incorporant elements heterogenis com és *cert que*, que han mostrat tenir un comportament clarament sistematitzable respecte de marcadors discursius estàndard (Alonso et al. 2004), però que tradicionalment són exclosos dels conjunts bàsics de marcadors que es sistematitzen en teories del discurs.
2. Per crear les taxonomies d'ETs per cada llengua, establim classes d'equivalència seguint la metodologia proposada per Hutchinson (2005) per l'anglès:
 - a) les classes d'equivalència s'estableixen mitjançant agrupació d'ETs similars, on la semblança entre ETs s'estableix mitjançant la divergència de Kullback-Leibler entre les distribucions de probabilitat de co-ocurrència de característiques rellevants amb cada ET.
 - b) les relacions de substituïbilitat s'estableixen mitjançant la variances en els valors d'expectació de co-ocurrència de característiques rellevants amb cada ET.Aquest mètode s'aplicarà de manera automàtica per cada llengua, i serà supervisat manualment. A partir de la supervisió manual s'interpretaran les relacions entre els diferents ETs, que seran etiquetats amb noms de relacions discursives clàssiques com "causa", "contrast", etc. En l'aplicació del mètode es prioritzarà la fiabilitat de les relacions trobades per sobre de la seva granularitat, de manera que es postularà una relació si i només si existeix evidència prou fiable. D'aquesta manera obtindrem un conjunt de relacions discursives altament fiable, encara que potser reduït.
3. Per unir les quatre taxonomies monolingües en una única taxonomia multilingüe, primer crearem taxonomies de parells de llengües i a partir d'aquestes crearem una taxonomia unificada. Per determinar la relació entre ETs de diferents llengües, aplicarem la metodologia de mapes semàntics amb jutges humans amb competència en les dues llengües implicades. L'objectiu dels experiments amb jutges és determinar la substituïbilitat dels parells de marcadors, com a mètode bàsic per determinar la seva relació en la taxonomia. Donat que les relacions de substituïbilitat solen ser imprecises, els grups de significat que podem determinar amb fiabilitat seran menys o igual de fins que en les taxonomies monolingües. Un cop establerta la taxonomia, s'interpretaran manualment les relacions entre ETs i se'ls donarà etiquetes discursives clàssiques, com "causa", "contrast", etc. Les etiquetes de la taxonomia multilingüe serà considerat el conjunt de significats discursius de referència per desenvolupar les aplicacions i recursos.
4. Utilitzant els ETs caracteritzats en 1. i la seva associació a una de les etiquetes de 3., desenvoluparem un analitzador discursiu. Es tractarà d'un analitzador parcial, ja que tan sols serà capaç d'analitzar aquells fragments de text que estiguin en l'abast d'un ET d'entre els que haurem tractat en 1. L'analitzador constarà de dos mòduls principals: un lèxic en què cada ET estarà associat a una de les etiquetes de 3., i una gramàtica en què es codificaran regles per manipular les diferents etiquetes, considerades com a diferents tipus semàntics. En la primera versió de l'analitzador, aquestes regles es crearan de manera totalment manual, en successives versions s'aplicaran mètodes d'inferència de gramàtiques a partir de corpus anotat per crear la gramàtica.

5. S'anotarà un petit corpus mitjançant l'analitzador automàtic. L'anàlisi automàtic serà millorat per jutges humans, que aplicaran el conjunt de relacions discursives obtingut en 3. Per garantir la consistència i reproduïbilitat de l'anotació de diferents jutges humans, es crearà un manual d'anotació.

Aquest corpus anotat semi-automàticament serà utilitzat per:

- 5.1. detectar nous ETs que aportin informació discursiva. Aquests ETs seran incorporados a la taxonomia descrita a 3., de manera que podran ser incorporats al mòdul lèxic de l'analitzador discursiu de 4.
- 5.2. induir una gramàtica d'anàlisi discursiu de corpus, que serà incorporada a l'analitzador discursiu presentat a 4.
- 5.3. avaluar el funcionament de l'analitzador.

Aquest pas serà repetit iterativament fins arribar a un sostre del mètode. En els subsegüents passos iteratius, el corpus a anotar serà diferent, de manera que s'incorporarà cada vegada un fragmento nou al corpus. Com és d'esperar que la precisió i la cobertura de l'analitzador augmentin en els diferents passos iteratius, esperem que l'esforç requerit per anotar corpus vagi disminuint progressivament, i, per tant, dedicant el mateix esforç en cada iteració, obtindrem quantitats de corpus progressivament més grans.

La difusió dels resultats obtinguts i dels recursos desenvolupats es durà a terme mitjançant la publicació d'articles en congressos i revistes. La difusió dels mètodes establerts es durà a terme mitjançant cursos i/o tallers especialitzats.

4. Pla de treball i calendari d'execució

Taula 1: distribució temporal de les tasques (1 mes = "o").

tasca	primer any				segon any			
	trim 1	trim 2	trim 3	trim 4	trim 1	trim 2	trim 3	trim 4
1.1. i 1.2. crear i analitzar un corpus d'ETs	o o o	o o o	o o o	o o o	o o o	o o o	o o o	o o o
1.3. identificar característiques contextuais de les ETs	o o o	o o o	o o o	o o o	o o o	o o o	o o o	o o o
2. crear i interpretar taxonomies d'ETs monolingües	o o o	o o o	o o o	o o o	o o o	o o o	o o o	o o o
3. crear i interpretar una taxonomia multilingüe	o o o	o o o	o o o	o o o	o o o	o o o	o o o	o o o
4. crear un analitzador discursiu	o o o	o o o	o o o	o o o	o o o	o o o	o o o	o o o
5. anotació de corpus	o o o	o o o	o o o	o o o	o o o	o o o	o o o	o o o
5.1. detecció de nous ETs	o o o	o o o	o o o	o o o	o o o	o o o	o o o	o o o
5.2. inducció de gramàtiques	o o o	o o o	o o o	o o o	o o o	o o o	o o o	o o o
5.3. avaluació de l'analitzador	o o o	o o o	o o o	o o o	o o o	o o o	o o o	o o o
6. difusió	o o o	o o o	o o o	o o o	o o o	o o o	o o o	o o o

Taula 2: durada absoluta de cada tasca.

tasca	durada
1.1. i 1.2. crear i analitzar un corpus d'ETs	2 mesos
1.3. identificar característiques contextuais de les ETs	8 mesos
2. crear i interpretar taxonomies d'ETs monolingües	2 mesos
3. crear i interpretar una taxonomia multilingüe	6 mesos
4. crear un analitzador discursiu	3 mesos
5. anotació de corpus	7 mesos
5.1. detecció de nous ETs	4 mesos
5.2. inducció de gramàtiques	4 mesos
5.3. avaluació de l'analitzador	4 mesos
6. difusió	8 mesos

Bibliografia

Alonso Alemany, L. (2005). Representing discourse for automatic text summarization via shallow NLP. tesi doctoral. Departament de Lingüística General, Universitat de Barcelona.

Carlson, L., D. Marcu i M. E. Okurowski (2003). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. Jan van Kuppevelt i Ronnie Smith (Eds.), Current Directions in Discourse and Dialogue. Kluwer Academic Publishers.

Freeling. <http://www.lsi.upc.es/~nlp/freeling>.

Hutchinson, B. (2005). The Automatic Acquisition of Knowledge about Discourse Connectives. PhD thesis. Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh, Regne Unit.

Knott, A. (1996). A Data-Driven Methodology for Motivating a Set of Coherence Relations. PhD thesis. School of Informatics, University of Edinburgh, Regne Unit.

Mann, W., i S. Thompson (1988). Rhetorical Structure Theory: Toward a functional theory of text organisation. Text, 3 (8), pp. 234 – 281.

Marcu, D. i A. Echihiabi (2002). An unsupervised approach to recognizing discourse relations. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), Philadelphia, Estats Units d'Amèrica.

The Potsdam Commentary Corpus. http://www.ling.uni-potsdam.de/cl/cl/res/forsch_pcc.en.html.

Reitter, D. (2003). Rhetorical analysis with rich-feature support vector models. Master's thesis. Universität Potsdam, Alemanya.

Wonsever, D. i J.L. Minel (2004). Contextual Rules for Text Analysis. Lecture Notes in Computer Science. Febrer 2004.