Un corrector i derivats

Martí Quixal
GLiCom, Universitat Pompeu Fabra

NLP Seminar, UPC
28 de novembre del 2008

Continguts

- Context
- Objectius
- Especificacions de correcció
- Recursos lingüístics
- Avaluacions
- Derivats
- Conclusions

Context

- La Generalitat de Catalunya vol fomentar la disponibilitat d'eines TIC per al català
  - SPL té interès en eines per fomentar l'ús (correcte i normatiu) del català (escrit)
  - STSI té interès en eines de codi obert
  - CTTI fa de gestor del projecte

Objectius (de projecte)

- Corrector ortogràfic i gramatical per al català:
  - Multiplataforma
  - Connectat amb programes ofimàtics més comuns (Office, OpenOffice, Mozilla, etc.)
  - Tractament de diverses variants dialectals, sempre en la versió normativa
  - Codi obert

Objectius (de grup)

- Crear una arquictectura de processament lingüístic general:
  - Multiplataforma
  - “Multilingüe”
  - Codi obert (si més no parcialment)
  - Modular i apta per a la recerca

Especificacions de correcció

- Tractament d'errors resultants en no paraules:
  - **casxa, *uqè*, **coneixo, *garantitzo*, etc.

- Tractament d'errors resultants en paraules:
  - **Lis va demanar permís*
  - *El que s'anomena *el síndrome d'Estocolm*
  - *Està orgullós *de que no hagis entrat a l'exèrcit*

(Les especificacions al complet:

http://parles.upf.es/corrector/EspecsErrorsElCorrector.pdf)

Recursos lingüístics

- Recursos lèxics
  - Lemari IEC 2006 (+ interferències)
  - Llistats d'abreviatures SPL
  - Llistats de topònims IEC + SPL
  - Gentilicis SPL
  - Antropònims
- Model de trigrames basat en CTILC

Recursos lingüístics (II)

- “Diccionaris” d’errors
  - Llistat de barbarismes freqüents (paraules soles i locucions/multiwords)
- Gramàtica de detecció d’errors basada en patrons morfosintàctics
  - Formalisme de mapping desenvolupat ad hoc
  - Regles manuals
- Corpus de frases anotades amb errors i proposta de correcció associada

Diccionari general

forma#lema:etiq-glic:dial:reg:freq-inv:caixa-estuct(#...)

- cap#cabre:VDR3S-:0:1:106333:LP
- cap#cabre:VRR2S-:0:1:51252463:LP
- cap#cap:EN--6S:0:1:1240:LP
- cap#cap:N5-MS:0:1:1530:LP
- cap#cap:P:0:1:2012:LP
- cap#capar:VDR1S-:3:1:0:LP
- cap#capítol:N5-MS:0:1:0:LD

- cantàssem#cantar:VJA1P-:3,4:1:0:LP
- cantam#cantar:VDR1P-:3:1:7321968:LP

Diccionari d'errors

forma#lema:etiq-glic:dial:reg:freq-inv:caixa-estuct(#...)

- adelantar#adelantar:VI----:0:6:10000000:LP

- boli#boli:N5-MS:0:3:10000000:LP
- bolis#boli:N5-MP:0:3:10000000:LP

- hooligan#hooligan:N5-6S:0:7:10000000:LP
- hooligans#hooligan:N5-6P:0:7:10000000:LP

<Style Id="6" ErrorCode="O_ERR_F_ALT_Z" Description="No admissible" UseForSuggestions="False" />
<Style Id="7" ErrorCode="" Description="Correcte-Incorrecte" UseForSuggestions="True" />

Corpus d’errors

El iai, el camí i {el hivern} seria un bon títol per a una pel·lícula xinesa.
@ Error{O_MAN_F_APR_E} Correction{l'hivern}

{La gat} menja peix.
@ Error{G_ERR_S_GEN_Z} Correction{el gat}

Fins a 3524 frases amb errors, codis i propostes de correcció.

Formalisme de detecció

- Les frases esdevenen un autòmata sense arcs d'ambigüitat
- Les regles s'apliquen com un chart amb look ahead
- Regles de mapping s'apliquen en funció de restriccions definides en un context

Formalisme de detecció (II)

<regla1 3 O_MAN_F_APR_E 1 2 1.lemma.EA--6S.F 2.form>

(ElForma NOT NULL)
(VI + Adj + Inv + Sg NOT NULL)
(Nom + Masc + Sg NOT NULL);

Avaluacions

- Avaluació de progrés (desenvolupament)
  - Corpus de desenvolupament
  - Incrementat progressivament amb exemples d'errors i algun exemple de frases sense error

- Avaluació manual (text desconegut)

Avaluació de progrés

Summary Statistics

Detection F 91% 3524 cs R 82% P 100% (2816tp, 94tn, 6fp, 608fn)

Classification Exact:	A 80%	3424 cs	E 20%	(2745p, 679f)
Classification Partial:	A 82%	3424 cs	E 18%	(2810p, 614f)
Location Exact:	A 70%	3382 cs	E 30%	(2384p, 998f)
Location Partial:	A 82%	3382 cs	E 18%	(2776p, 606f)
Correction Exact:	A 57%	3415 cs	E 43%	(1949p, 1466f)
Correction Partial:	A 63%	3415 cs	E 37%	(2146p, 1269f)

(894 multierror cases)

Avaluació de progrés (II)

Detection: False Positives by Error Proposed (4 error codes)

T_SOB_F_ALT_E: 50% (3 cs)
O_MAN_F_APR_P: 17% (1 cs)
(…)

Classification: Accuracy Values by Error Expected (47 error codes)

T_ERR_F_PTV_Z	A 100%	06 cs	E 00%	(6p, 0f)
O_MAN_F_ACC_D	A 80%	44 cs	E 20%	(35p, 9f)
G_ERR_S_GEN_Z	A 80%	1310 cs	E 20%	(1054p, 256f)
G_ERR_S_MOD_C	A 67%	03 cs	E 33%	(2p, 1f)

(…)

(I diverses informacions per al debugging i logging)

Avaluació manual

Error detection

Engine	Prec	Rec
EC	70%	83%
MG	56%	78%

Proposal generation

Engine	Correct	Inadeq	None	Irrelev
EC	55.8%	20%	7.4%	16.8%
MG	46.7%	24%	2.7%	26.6%

Manual evaluation of EC/MG on a small corpus (6 texts, 2300 words) from a variety of genres (newspaper articles, blog text, high school and college essays and informal e-mails).

Anàlisi d'errors (undetailed)

Causes dels errors no detectats:

- Ambigüitat: "ho fa tot l'afició". Vol dir "tota l'afició ho fa" o "l'afició ho fa tot"?
- Errors de tagging "Totes dues modalitats nous són vàlides" ("nous" rep lectura verbal)
- Regles evitades per reduir soroll:
  - (1) *Un gran nombre de cotxes barat a la venda
  - (2) És un antic company d'escola presentat per en Joan
- Algunes regles mal definides també…

Anàlisi d'errors (undetailed, II)

Causes d'errors en propostes de generació:

- Limitacions del formalisme a l'hora de generar propostes de correcció:
  - "El nedar no m'agrada" no genera "Nedar"
  - "L'armari blaus…" genera "L' armari blau"
  - "Ho fan tan alegrament" → no genera "alegrement"

Derivats

- LINLaP: tagger (CA i ES)
- COTiG: motor de revisió i correcció (CA)
- Corpus d’errors del català

Derivats (II)

- Diccionari(s)
  - General
  - D’errors (triggers i generació d’alternatives)
  - Especials:
    - Abreviatures
    - Topònims
    - Antropònims

Aviat disponible a lafarga.cat (tagger, no)

L’equip

- Programadors:
  - Francesc Benavent
  - Beto Boullosa
  - Daniel Chicharro
  - Bernat Grau
  - Marc González
  - Joan Moratinos
  - Oriol Valentín
- Lingüistes:
  - Judith Domingo
  - Benito Hellín
  - Guillem Massó
  - Òscar Puente

Coordinador: Martí Quixal

Director: Toni Badia

El que hem après

- La performance en general és bona, però hi ha mòduls millorables (tokenitzador, hmm, etc.)
- El temps de resposta ha estat crític, igual que el consum de memòria
- Per a aplicacions de llenguatges controlats, l'eina és força flexible. També per d'altres amb una quantitat raonable de millores

Més coses apreses

- El sistema de regles basades en patrons morfosintàctics (forma+lema+POS) va prou bé, però:
  - Les regles s'han de restringir molt per evitar falsos positius
  - Es multipliquen enormement per tractar errors de concordança
- Ha costat molt més de fer el que envolta el motor que el propi motor

Gràcies per la vostra atenció!

http://parles.upf.es/corrector