Morphology to the Rescue Redux:
Resolving Borrowings and Code-mixing in Machine Translation

Esmé Manandise
IBM Research

Claudia Gdaniec
South Westphalia University of Applied Sciences

SFCM-2011
August 26, 2011, Zurich, Switzerland

Preface

- Slot Grammar analysis and LMT machine translation for Spanish-English has been funded since 2003 by IBM Corporate Citizenship & Corporate Affairs (CCCA).
  - CCCA .overlooks programs of corporate social responsibility and deals with social issues confronting Hispanics in the U.S (Latinos and IT; Latino youth school achievement).
  - The translation portal (www.traduceloahora.org), which has been operational since Aug. 2003, includes hundreds of non-profit organizations and schools. It offers free access to Spanish-English translation of (1) Web pages, (2) Web searches , (3) Emails, and (4) Instant chats

- **LMT Team**
  - **Current Members**

    Nelson Correa, LMT English=>Spanish, CCCA liaison, WTS and website operation

    Esmé Manandise, LMT French⇔English and Spanish =>English, Romance Languages SG parser, email preprocessor, email MT, cross-lingual search

    Michael McCord, Team leader for Language Analysis and Translation, SG/LMT shell, ESG, DeepQA and Watson

  - **Associates**

    Arendse Bernth, EasyEnglishAnalyzer and TermExt

    Claudia Gdaniec, LMT German⇔English

Main Points

- Borrowings and code-mixing with zero-morph are similar to neologisms, i.e. words unfound in Spanish, the source language. For example, borrowing "*startear*" from "*start+ear*" or code-mixing "*Anyway deje las keys **adentro**"* (*Anyway I left the keys inside*)"
- Analysis of unfound words through a set of derivational morphological rules assigns these words morphological, syntactic, and semantic features
- Language-independent strategy
- Utility to acquire these unfound words in an LMT-compatible format in an auxiliary bilingual lexical file subsequently merged into the core lexicon
- Output generation of unfound words through a set of derivational morphological rules in the transformational component
- Use of semantic and syntactic features for both analysis and transfer generation

More Specific Claims

- Rule-based approach to phenomena of languages in contact like borrowing benefits from linguistic generalizations needed to analyze an individual language. This approach relies on existing rules that are needed by the morphological analyzer independently of borrowing and code-mixing
- Great variability and unpredictable frequency of occurrences of specific borrowings and code-mixing instances make corpus-based analyses difficult
- Basic features and general principles of the LMT morphological analyzer have the language-independent flexibility needed to support the creation of rules for the analysis of words with affixes from Spanish and the lexical bases of English
- Strategy of using an existing morphological analysis developed for one language and extend it to handle word formation in the context of two (known) languages in contact

Fundamentals of Email Processing

- No restrictions on the content or quality of the input
- Translated output must be intelligible to recipients
- “Garbage in, garbage out” is not acceptable
- Processing input requires reducing input “noise”, that is, cleaning input prior to full analysis of input

Spanish Input Emails

Most Frequent, Non-trivial Issues

- Lack of punctuation
- Lack of written diacritics (accents)
- Misspelled words
- Borrowings from English
- Code-mixing of Spanish and English
- Code-switching between Spanish and English
- Variants of Spanish
- Poor syntax

Complexity of the Task

Examples

Email Example (1)
Monolingual Spanish-speaking Parent

mastra nesesito saver como esta mi ijo en la clase grasias

Standard/Prescriptive Spanish

Maestra, necesito saber cómo está mi hijo en la clase. Gracias.

Standard/Prescriptive English

Teacher, I need to know how my son is doing in class. Thanks.

Words: 11

Deviations from SPS: 7

Email Example (2)
Predominantly Spanish-speaking Student in English-speaking School

thanks mrs. griffin usted asido muy buena con migo y con mis companeros me da mucho help le boy a *echar ganas* para poder aprender ingles.

Standard/Prescriptive Spanish

Gracias, Sra. Griffin. Usted ha sido muy buena conmigo y con mis compañeros. Me ayuda mucho. Le voy a dar ganas para poder aprender inglés (?).

Standard/Prescriptive English

Thanks, Mrs. Griffin. You have been very nice with me and my buddies. You help me a lot. I am going to become motivated to learn English.

or

I am going to motivate you so as to be able to teach me English.

Words: 26

Deviations from SPS: 12

Email Example (3)
School Administrator Using Spanish

Los padres ahora pueden participar y ser envueltos en las **policy makers** que en el pasado. Esta participatión trae a los padres de los niños a las escuelas hacia el proceso de **hacer** las desiciones, que también son llamados **site-based management**, este proceso envuelve a los maestros y padres y a los administradores, **counselores**, **principales**, y ***superintendentes***. Juntos determinan las necesidades individuales de la escuela y **hacen** decisiones y hacen plañes que creen que haran la escuela ***efectiva***.

Standard/Prescriptive Spanish

Los padres ahora pueden participar y ser envueltos en las formulaciones de política como en el pasado. Esta participación ayuda a los padres de los niños de las escuelas a tomar las decisiones, que también son llamadas gestión local. Este proceso envuelve a los maestros y a los padres y a los administradores, consejeros, directores de escuela y de districto escolar. Juntos determinan las necesidades individuales de la escuela y toman decisiones y hacen planes que creen que haran la escuela más eficaz.

Standard/Prescriptive English

Parents can now participate and be involved in policy formulations as in the past. This participation helps parents of children in schools to make decisions, which are also called local management. This process involves teachers and parents also administrators, counselors, principals and superintendents. Together, they determine the individual school needs and make decisions and make plans they believe will make the school run more effectively.

Words: 77
Deviations from SPS: 16

Cityspeak, Fictive Cousin of Spanglish?

When recollecting his meeting with Detective Gaff, Rick Deckard in *Blade Runner* says,

"That gibberish he talked was city-speak, guttertalk, a mishmash of Japanese, Spanish, German, English, what have you. I didn't really need a translator. I knew the lingo."

Likewise, LMT Spanish-English must learn to handle the lingo (Spanglish).

Example (4)
The Jewel in the Crown
Fully-fleshed Spanglish (E. Gonzales)

Bro: Fue one día like today, hace años, when I first llegué a La Yuma. It wasn't el hielo sino Miami, antes called "The capital of the Cuban exilio" pretty parecida a Cubita la bella, hot and humid, con el español and spanglish que se escuchaba all around, sobre todo en la Sauesera. Al día siguiente I went to la Migra to meter papeles y apply por un estato legal. In the meantime, ya empezaba to work doing patios, o sea, cutting grass en las yardas de Miami. Un amigo was already working de junkero y un relative de friends in Cuba ***bregaba*** en una pompa. **Pa'** tener a good job I had to moverme pero didn't have plata for a buen ***carro***, so le compré un ***tranporteichon*** a una ***yoni***. It lasted about un año, ***yompeándolo*** algunas veces, of course.
I shared a duplex con my familia but other amigos I knew rentaron un ***efiche***.
They had to pagar casi the same and had solo one cuarto!
Después I got mi tarjeta con el número del Social and a few semanas later pude landear un better job. Tenía que trabajar no less than 12 hours por día, but I felt bien, **'coz** tenía libertad **pa'cer** whatever yo quisiera. I incluso worked en un funeral home y as a teacher!
Tuve friends que used to work as pomperos, ruferos, yunkeros, grueros, en tormotos, de **dílers**, serving mesas, carpeteros, en los desks de los hoteles, como reps, troqueros, vendiendo **áiscrim** y balloons, but always estudiando y **mechándose pa'** salir ahead.
Nada de ***janguear*** con la wrong ganga.

Dealing with the Input Noise

(A) Preprocessing Tasks

versus

(B) Morphological Analysis

(A) Preprocessing/Cleaning before Submission to LMT for Analysis

- Accent restoration
- Some types of misspelling
- Expressive repetition of letters and symbols
- Tagging of proper names
- Tagging of affiliations (department, company…)
- Tagging of addresses (street, city, country…)

Stand-alone program (mpro) written in C

(B) Morphological Analysis of Input

- Use morphological analyzer of Standard Spanish (written in C) for derivational and inflectional affixes
- Treat borrowings (established, infrequent, one-time occurrence) as any word of the Spanish language that can be subjected to word formation analysis
- Modification as to where to look for the lexical base for affixation
- Expand the lexicon with borrowings

Analysis of Unfound Words in LMT

- Use strategy developed for French and German
- Expand it to Spanish
- Borrowings as unfound words
- Code-mixing as special case of unfound words

Unfound Words

- A lexical base in the source lexicon
- Morphologically-related words listed in the source lexicon
- A dummy base
- Rules applied are those of the source language without reference to the target language

Unfound Words for Languages in Contact

- Rules applied are those of the source language **with** reference to the target language lexicon (lexical base and morpho-syntactic and semantic features associated with that lexical base)
- By assuming similarities in morphological processes and rules between affixes applied to borrowings found in emails and other source-language affixes that have not (yet) been found in email, the scope of the analysis expands
- Increased coverage of borrowings validates analysis

Morphological Analyzer: Main Steps

- Affix stripping and base spelling adjustments and stem changes
- Lexical lookup
- Affix operations

Analysis of Borrowings

- Isolate a base
- Isolate affixes
- Create complex, bracketed word structures
- Determine a part of speech together with morpho-syntactic features and, where possible, semantic and syntactic features
- Flag the word as unfound and borrowed

The Transfer of Borrowings

- Create a target-consistent string or subtree
- Integrate the transfer and its modifers correctly into the target tree

Language-independent Strategy Unfound Words in LMT French

(5) ([affirmer + tif])
adjective masculine singular
"attribute")

case s_tif:
if a word has been found without derivational affixes,
then fail;
else if the POS is not verb,
then fail;
else
newPOS = adjective;
gramm_gender = masculine;
newSemantictype = "action";
newWordstructure = wordstructure + tif;

(6) ([dummyverb + tif])
adjective masculine singular
"attribute")

if word is dummyverb,
penalty = 10;
else
penalty = oldpenalty+1;

if inflectional affix in
affix list,
do not set morphosyntactic
features here;
else if no other affixes in
affix list,
morphosyntax = masculine singular

Language-independent Strategy Unfound Words in LMT French

(7)([[affirmer + tif] + ment]
adverb "manner")

(8) Transformation Derived.3 on node #, "assert", produces tree
|__adv assert ([POS adv] [Sslot
vadv][SSem “manner”])

(9) Transformation Manner_adv.1 on node #, "asserting", produces tree...
|_____vadv **in** ([POS prep])
| __ndet **a** ([POS det] [Number sg] [Person pers3])
| |__nadj **assert** ([VInfl ving] [POS verb])
| |
|__|__objprep **manner**([POS noun] [Number sg] [Person pers3])

The final English output for *affirmativement* is *in an asserting manner*.

French Example

(10) **Input:**

Radmirons l' *admirable admirateur admiratif* et l'*admirable admiratrice admirative*.

Old result:

Radmirons admirable admirateur admiratif and admirable admiratrice admirative.

New result:

Radmirons the admirable admiring admirer and the admirable admiring admirer.

Language-Independent Strategy Spanish Example

(11) **tranquilamente**

Bases and known affixes:
tranquilamente
tranquilo:mente
dummy:mente

Resulting Word Structure:

([tranquilo + mente] adverb suffixed manner)

Language-Independent Strategy Spanish Example

(12) **startear**

Bases and known affixes:

a. startear

b. start:ear

c. starte:ar

d. dummy:ear

Resulting Word Structure:

([start + ear] verb infinitive suffixed (penalty 10) borrowed/notfnd)

Language-Independent Strategy Spanish Example

(13) **Anticooldad**

Bases and known affixes:
a. anticooldad
b. anticool: dad
c. cooldad : anti
d. cool: anti, dad
e. dummy: anti, dad

Resulting Word Structure:

([anti+cool+dad] noun f sg property attribute prefixed suffixed (penalty 10) borrowed/notfnd)

Parse without Borrowing Strategy

(14) Miguel fue a startear su carro, pero no arrancó.

Incomplete parse:
Syntactic analysis no. 1 Evaluation = 0.750000...

```
---------------------------------------------------------------------
o----- top          incomplete(10)    incomplete
| .--- subj(n)      miguel1(1)        noun propn sg m h
`-+--- u            ir1(2,1,u,3)      verb vfin vpast vsg vpers3
| `--- comp(p)      a1(3,4)           prep pprefv motionp
| `--- objprep(n)   startear(4)       noun propn sg (notfnd)
| .--- ndet         su1(5)            det sg m possdet ingdet
`----- u            carro1(6)         noun cn sg m (st_vehicle m)
`----- u            pero1(7)          conj
| .--- vadv         no1(8)            adv ppadv neg
`----- u            arrancar1(9,u,u)  verb vfin vpast vsg vpers3 vind
---------------------------------------------------------------------
```

Miguel went to startear his car but he did not start.

Parse with Borrowing Strategy

(15) Miguel fue a startear su carro, pero no arrancó.

Syntactic analysis no. 1 Evaluation = 17.531110....

```
-----------------------------------------------------------------------
.--------- subj(n)        miguel1(1)    noun propn sg m h
.-+------- lconj          ir1(2,1,3,u)  verb vfin vpast vsg vpers3
| `------- comp(pinf)     a1(3,4)       prep pprefv infobj motionp
| `------- objprep(binf)  (sfx startear)(4,u,6)
                                        verb vinf
                                        borrowed/notfnd
| | .----- ndet           su1(5)        det sg m possdet
| `------- obj(n)         carro1(6)     noun cn sg m (st_vehicle)
o--------- top            pero1(7)      verb vfin vpres vpast vsg
| .------- vadv           no1(8)        adv ppadv neg
`--------- rconj          arrancar1(9,u,u) verb vfin vpres vpast vsg
-----------------------------------------------------------------------
```

Miguel went to start his car but it did not start.

Code-mixing

- Analysis changed since paper submission
- Direction of current development:
  1. Run Spanish morphological analyzer and lexicon
  2. Run unfound words with English morphological analyzer and lexicon
  3. Make output of 1 and 2 available to Spanish parser

Example Parse without Code-mixing Handling

Here la mujer added imagenes.

```
Incomplete parse:

Syntactic analysis no. 1   Evaluation = 0.150000...
-----------------------------------------------------------------------
   --
o--- top   incomplete(6)    incomplete
`--- u     Here(1)          noun propn sg (notfnd)
| .- ndet  la1(2)           det sg f def
`-+- u     mujer1(3)        noun cn sg f humind st_anim
| `- nprop added(4)         noun propn sg (notfnd)
`--- u     imágenes1(5,u,u) noun cn pl f st_communication st_art
-----------------------------------------------------------------------
Here the woman added images.
```

Example Parse with Code-mixing Handling

Here la mujer added imagenes.

```
Syntactic analysis no. 1   Evaluation = -2.190000...
-------------------------------------------------------------
.--- vadv    here1(1)          adv nocompare introadv loc
                               (code-switching/notfnd)
| .- ndet    la1(2)            det sg f def
.--- subj(n) mujer1(3)         noun cn sg f h humind st_anim
o--- top     added1(4,3,5,u)   verb vfin vpast vsg vpers3 vind
                               vsbuj (code-switching/notfnd)
`--- obj(n)  imágenes1(5,u,u)  noun cn pl f (st_communication
                               st_art
-------------------------------------------------------------
   Here the woman added images.
```

Conclusions

- Results
  - Better parsing
  - Better output; in most cases, the automatically derived transfers are acceptable

    startear → start

    anticooldad → anticoolness
  - Mechanism for generating or accessing transfers is flexible because it is performed in the transformations with access to whole subtrees in the target, which can avoid word-by-word translation

- Drawbacks
  - Less efficiency of processing time and space
  - Sometimes transfers may not be easily understandable

Appendix: Europarl Corpus

- Spanish→English (sample of 660,050 words, **less than 0.0005% of instances of borrowings, neologisms, code-mixing**)
- Run a LMT utility for Europarl words unfound in LMT SE lexicon
  - Proper names, affiliations, product names, organizations, etc… like Moreira, Duhamel, Eirecom, IFOP, “Norwegian People ' s Aid”
  - Misspelled words like in “prin**i**cipos” → “principios”, or “osbtante” → “obstante
  - Spanish neologisms like “vinculatoriedad”
  - Borrowing like “prioritización”, “similaridad”
  - Code-mixing (with variants) like “los left overs”, “los leftover” , “el cuarto left over“

Appendix: Spanish examples in Europarl

- Few instances of identical non-Spanish words
- Maximum of instances in Europarl sample is 4

  1 instance:

  ¿ Es que estos países pueden preciarse de garantizar lo que en inglés se denomina **good governance**, buen gobierno?

  2 instances:

  Así pues si esta agencia no alcanza los **standards** que puedan convertirla en un organismo capaz de imponer una política , no creo que pueda hacer grandes cosas .

  Dichos procedimientos ya han sido modificados y cambiados y ahora contamos con nuestra propia " Food **Standards** Agency“.

Contact Information

Esmé Manandise, IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598, USA, esme@us.ibm.com, esmeman@comcast.net

Claudia Gdaniec, South Westphalia University of Applied Sciences, 59494 Soest, Germany, gdaniec@fh-swf.de