A Comparison of Letter-Level Classifiers for Portuguese Diacritic and Capitalization Restoration

Brian Rosenfeld

Introduction

- Classifiers for restoring diacritics and capitalization to Portuguese text at the letter level
- Compare the performance of different “window” sizes (N = 0, 1, 2, 3)
- Compare the performance of Naïve Bayes and Decision Tree classifiers with a baseline
- Explore rules for Portuguese diacritics and capitalization

Introduction

Isso é um caso de polícia. No início, eu sofri ameaças, ligavam para a minha casa, ameaçaram a minha esposa e tive que trocar todos os telefones. Mas essas pessoas não são torcedores. Elas se aproveitavam do Cruzeiro para ganhar dinheiro.

This is a police matter. At first, I suffered threats, linked to my home, threatened my wife and I had to replace all phones. But these people are not fans. They took advantage of the cruise for earning money.

isso e um caso de polícia. no início, eu sofri ameacas, ligavam para a minha casa, ameacaram a minha esposa e tive que trocar todos os telefones. mas essas pessoas nao sao torcedores. elas se aproveitavam do cruzeiro para ganhar dinheiro.

That and a case of police. at first, I suffered threats, linked to my home, threatened my wife and I had to replace all phones. but these people are not fans. they took advantage of the cruise for earning money.

Motivation

- Proofing tools
  - Ex. Microsoft Word spell check
- Auto-correction when typing Portuguese on an English keyboard
- Preprocessing for other NLP tools
  - Adding diacritics and capitalization to transcribed speech or 7-bit ASCII
  - Allows for tagging, parsing, translation, etc.

Yarowsky (1999)

- Corpus-based techniques for accent restoration
- N-gram tagger with suffix annotations and function words
  - ex. *la/LA posicion/-IÓN anuncio/-Ó oficialmente/-MENTE*
  - Ignores lexical associations (ex. subjunctive)
- Naïve Bayes
  - Outperformed N-gram on context sensitive distinctions
- Decision Lists
  - Combines strengths of N-gram and Naïve Bayes

Milhacea (2002)

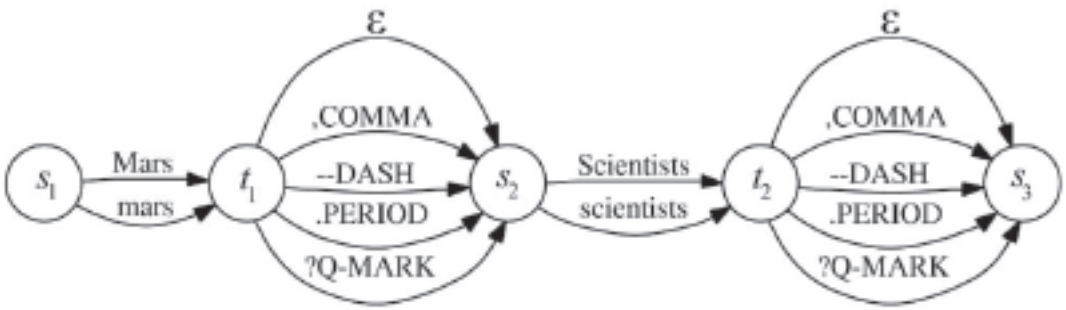
- Diacritic restoration at the letter level
- Used an instance based learning algorithm and a decision tree classifier
  - Similar performance between models
  - Best window size was 10 letters (N = 5)
  - Reported accuracy of over 99% and better speed than word-level algorithms

Chelba and Acero (2004)

- Capitalization as tagging problem
- Maximum entropy Markov model (MEMM) and maximum a-posteriori (MAP) adaption technique
- MEMM performed 35-45% better than 1-gram model
- Adding a small amount of domain-specific data reduces error over baseline by 50-60%

Gravano, Jansche, Bacchiani (2009)

- Single-pass n-gram model for both capitalization and punctuation
- Majority of capitalization errors were due to punctuation errors



Batista, Trancoco, Mamede (2012)

- Recovery of capitalization and punctuation in Portuguese and English speech transcripts
- Two classification tasks with logistic regression (maximum entropy) classification models
  - Capitalization used first; helps create part of speech tagger
  - Part of speech tagger used to help with punctuation
- ME model outperformed Hidden Markov Model for speech but not written text

Method: Preprocess Data

- Python scripts for preparing arff files
  - Use unidecode module for converting to ASCII
  - Convert each character to lower-case, base-7 ASCII and lookup its accent code

Accent Mark	Lower	Upper
NONE	0	6
`	1	7
´	2	8
^	3	9
~	4	10
ç	5	11

- ARFF format
  - @RELATION portuguese
  - @ATTRIBUTE letter {a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, SPACE, PERIOD, COMMA, OTHER, BOS, EOS}
  - @ATTRIBUTE target {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}

Method: Train Classifier

- Weka 3: Data Mining Software in Java
  - NaiveBayes and J48 classifiers with cross-validation and 10 folds
  - Trained classifiers for different windows (N=0 [baseline], N=1, N=2, N=3)
- Why these classifiers?
- Expectations?

Assumptions

- Divided non-letters and numbers into SPACE, COMMA, PERIOD, OTHER
- Mixed Brazilian and European Portuguese
  - In some cases, two different classifications should be correct
  - European: *cómoda* vs. Brazilian: *cômoda*

Data

- Subset of Portuguese Wikipedia dump
- 551,496 characters (83,610 words)
- Assessment:
  - Used outside program for parsing text from xml
    - Should be fairly accurate
  - Brazilian vs. European Portuguese
    - Might cause some inconsistency
  - Noisy? Foreign words and pronunciation guides

Results

- Baseline: 94.87%
- Decision Tree outperformed baseline and Naïve Bayes
- Decision Tree performance improves in increasing window size whereas Naïve Bayes worsens in increasing window size

Window	Naïve Bayes	Decision Tree
1	95.19%	96.24%
2	93.90%	97.61%
3	93.27%	97.90%

Results

a	b	c	d	e	f	g	h	i	j	k	l	
484442	702	5600	996	520	791	11821	1108	353	2	0	0	a
202	65	0	0	0	0	26	0	0	0	0	0	b
4526	0	2520	38	2	0	20	0	5	0	0	0	c
404	0	92	761	0	0	7	0	0	0	0	0	d
401	0	1	1	3499	0	0	0	0	0	0	0	e
424	0	0	0	0	2260	0	0	0	0	0	0	f
6048	9	7	0	3	15	5924	0	2	0	0	0	g
2	0	0	0	0	0	4	0	0	0	0	0	h
45	1	6	0	0	0	54	0	2	0	0	0	i
1	0	0	0	0	0	1	0	0	0	0	0	j
0	0	0	0	0	0	0	0	0	0	0	0	k
0	0	0	0	0	0	0	0	0	0	0	0	l

Naïve Bayes (false positives)

49% accuracy for accent-free, upper-case letters

a	b	c	d	e	f	g	h	i	j	k	l	
504980	0	564	75	28	54	625	0	9	0	0	0	a
291	0	0	0	0	0	2	0	0	0	0	0	b
2684	0	4392	27	1	0	7	0	0	0	0	0	c
487	0	10	767	0	0	0	0	0	0	0	0	d
225	0	0	1	3676	0	0	0	0	0	0	0	e
122	0	0	0	0	2560	2	0	0	0	0	0	f
5793	0	2	0	0	8	6200	0	5	0	0	0	g
1	0	0	0	0	0	5	0	0	0	0	0	h
56	0	2	0	0	0	4	0	46	0	0	0	i
2	0	0	0	0	0	0	0	0	0	0	0	j
0	0	0	0	0	0	0	0	0	0	0	0	k
0	0	0	0	0	0	0	0	0	0	0	0	l

Decision Tree (false negatives)

51% accuracy for accent-free, upper-case letters

Results

- Sample paths from the decision tree N = 1
  - prev = c, letter = a, next = o → target = 4 (~)
    - 1533 correct (99.67%); 5 incorrect
    - *ção* is a common noun ending
  - prev = SPACE, letter = e, next = SPACE → target = 0 (none)
    - 3109 correct (83.96%); 594 incorrect
    - *e* means "and" while *é* is 3rd person singular form of "to be"
  - prev = a, letter = c, next = a → target = 5 (ç)
    - 1141 correct (89.84%); 129 incorrect
    - *maça*, *raça destacar, acabar*

Results

- No direct comparison to past results
- Diacritics
  - Milhacea (2002) achieved over 99% for diacritics (ex. *i* vs î) at the letter level but on Romanian and with more data
  - Similar to Yarowsky (1999) with different performance for different diacritics
- Capitalization
  - On word level, Chelba and Acero (2004) had 2.3% error with maximum entropy model

Future Work

- Further increasing the window size for a decision tree model
- Differences between diacritics for European and Brazilian Portuguese
- A maximum entropy classifier with this approach
- Two-pass algorithm with accents on the character level and capitalization at the word level
- Effect of removing punctuation and spaces and using word barriers instead

Conclusions

- Decision Tree classifier better suited than Naïve Bayes classifier and baseline (1-gram)
- High diacritic accuracy can be achieved for certain kinds of accents (ex. nasal sounds)
- Relatively poor performance on capitalization
- Realistic method for languages with limited resources

References

1. Batista, F., Moniz, H., Trancoso, I., & Mamede, N. (2012). Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. *Audio, Speech, and Language Processing, IEEE Transactions on*, *20*(2), 474-485.

2. Chelba, C., & Acero, A. (2006). Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, *20*(4), 382-399.

3. Gravano, A., Jansche, M., & Bacchiani, M. (2009, April). Restoring punctuation and capitalization in transcribed speech. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* (pp. 4741-4744). IEEE.

4. Mihalcea, R. F. (2002). Diacritics restoration: Learning from letters versus learning from words. In *Computational Linguistics and Intelligent Text Processing* (pp. 339-348). Springer Berlin Heidelberg.

5. Yarowsky, D. (1999). A comparison of corpus-based techniques for restoring accents in Spanish and French text. In *Natural language processing using very large corpora* (pp. 99-120). Springer Netherlands.