Detecção de Variação de Grafia no Córpus Milênio

Rafael Giusti
Orientadora: Prfª Drª Sandra Maria Aluísio

Núcleo Interinstitucional de Lingüística Computacional – NILC
Instituto de Ciências Matemáticas e de Computação – ICMC
Universidade de São Paulo – USP

Tópicos

- Propósito
- Abordagem
- Regras de transformação
- Expressões regulares (?)
- Aplicação
- Processo
- Relatórios
- Regras em uso
- Objetivos em mente

Propósito

- Ausência de ortografia produz variantes gráficas de uma mesma palavra
  - muito, muyto
  - “Regra” no português histórico
  - Freqüente no córpus Milênio

Propósito

- Variação de gráfica é problemática para contagem de freqüência
  - Verifica-se 30 ocorrências de “muito”
  - Verifica-se 23 ocorrências de “muyto”
  - Verifica-se 60 ocorrências de “pedra”

Quais verbetes serão selecionados para o nosso dicionário?

Propósito

- Solução:
  - Verificar automaticamente a variação de grafia
    - Reconhecer que as grafias “muito” e “muyto” estão associadas a uma mesma idéia
    - Efetuar contagem de tais grafias
    - Reconhecer que a palavra “muito” possui 53 ocorrências no córpus, mesmo com tais irregularidades

Abordagem

- Várias abordagens foram pensadas:
  - Agrupamento por distância de edição
    - Quantidade de passos de inserção, exclusão ou troca de símbolos
      - ca**s**a e ca**z**a (1)
      - al**f**a**c**e e al**ff**a**ss**e (3)
    - Propenso a erros
      - gra**n**de e gra**d**de (1)
      - grad**de** e grad**i** (2)
    - Computacionalmente caro
      - Sem um dicionário prévio, seria necessário comparar todas as grafias

Abordagem

- Agrupamento por análise fonética
  - Grafias são convertidas para sua equivalente fonética
  - As equivalentes são comparadas por distância fonética
  - Problemas:
    - Extremamente caro computacionalmente
    - Propenso a erros
    - Ausência de ortografia não permite abstrair um relacionamento lógico entre as grafias e a pronúncia

Abordagem

- Regras de normalização
  - Dissertação de mestrado de Alexandre Hirohashi
  - Regras de normalização
  - Transformação de grafia em textos do córpus Tycho Brahe
  - Avaliação do aprendizado por comparação com um córpus manualmente anotado
  - Aprendizado completamente automático
  - Problemas
    - Não temos um córpus manualmente anotado

Abordagem

- Abordagem escolhida
  - Aprendizado por regras de transformação
    - Regras criadas manualmente
    - Detecção da variação de grafia é completamente automatizada

Regras de transformação

- Três partes: um contexto de condição, um contexto de substituição e uma seqüência de substituição
  - Contexto de condição: usado para escolher as palavras que são afetadas pela regra
  - Contexto de substituição: usado para escolher o trecho da palavra que será alterado
  - Seqüência de substituição: substitui o contexto de substituição

Regras de transformação

- Exemplo
  - np n m
    - Condição: np
      - se**np**re
      - co**np**anhia
    - Substituição: n
      - Seleciona somente o n de np
    - Seqüência: m
      - Troca o "n" selecionado por "m"
        - senpre → sempre
        - conpanhia → companhia

Regras de transformação

- As três partes de uma regra de transformação são expressões regulares
  - Permite maior flexibilidade na definição dos contextos
    - n[pb] n m
      - se**np**re
      - co**np**anhia
      - deze**nb**ro
      - e**nb**argo

Expressões regulares

- Cadeias de símbolos que
  - Seguem uma norma sintática
  - Representam ou cobrem um conjunto de cadeias de símbolos
- Finalidades
  - Descrever linguagens
    - Por exemplo, grafias "incorretas"
  - Casar padrões
    - Por exemplo, detectar uma variação de grafia

Expressões regulares

- Casamento de padrão
  - Ocorre se existe uma correspondência entre uma subcadeia do texto e toda a cadeia da expressão regular
    - A expressão bola casa com as cadeias `bola`cha, `bola` e `bola`do, mas não com os padrões `cola`, `bolo` e `ola`.
    - A expressão alvo casa com as cadeias s`alvo`, e `alvo`rada, mas não com `salvaguarda`.

Expressões regulares

- Símbolos especiais permitem casamento de padrão avançado
  - `.` (ponto)
  - `[X]` (subconjunto)
  - `$` (fim-de-cadeia)
  - `^` (início-de-cadeia)
  - `?` (opção)

Expressões regulares

- Símbolo `.` (ponto)
  - Casa com qualquer símbolo não-especial do alfabeto das expressões regulares
    - `bol.` casa com `bola` e `bolo`
    - `dezo.to` casa com `dezoito`, `dezoyto`, `dezojto` e `dezouto`
    - `...` casa com qualquer seqüência de três letras, incluindo palavras e anagramas

Expressões regulares

- Subconjunto
  - É um pouco mais restritivo do que o ponto (`.`). Este símbolo define um subconjunto de símbolos que podem casar com a cadeia
    - `n[pb]` casa com `te`<u>`np`</u>`o` e `le`<u>`nb`</u>`rança`, mas não com tempo e lembrança
    - `[ao]` casa com qualquer artigo singular, não com ao
    - `ch[aã]o` casa com <u>`chão`</u> e `col`<u>`chao`</u>, mas não com `cham` ou `chãm`

Expressões regulares

- Fim-de-cadeia ($)
  - É uma "âncora"
  - Obriga com que aquele trecho do padrão seja casado com o final da cadeia
    - `capa$` casa com `capa` e `contracapa`, mas não com `capacete`
    - `[aei]r$` casa com todas as palavras finalizadas em `ar`, `er` ou `ir` (verbos)

Expressões regulares

- Início-de-cadeia (^)
  - Âncora
  - Força o casamento de padrão a acontecer no início da cadeia
    - ^capa casa com capa e capacete, mas não com contracapa
    - ^auto casa com todas as palavras que tem prefixo "auto" (auto-confiança, auto-motivação) ou que apenas começam com a cadeia "auto" (autoridade, autorama)

Expressões regulares

- Símbolo ? (opção)
  - Casa com a cadeia mesmo que o símbolo anterior não case
    - `casc?a` casa tanto com `casa` quanto com `casca`, pois o símbolo "c" se torna opcional
    - `casc?a` não casa com `caca`, pois o símbolo "s" não é opcional
    - `[ao]s?` casa com todos os artigos definidos a, o, as, os, pois o símbolo "s" é opcional

Aplicação

- Hirohashi utiliza regras de transformação para normalizar textos históricos
- Nós utilizamos regras de transformação para agrupar grafias
  - Se duas grafias originais O_1 e O_2 são transformadas para uma mesma grafia resultate R, então O_1 e O_2 se agrupam em torno da mesma palavra
    - Variação de grafia para uma palavra qualquer

Aplicação

- Exemplo:
  - As seguintes grafias são encontradas no córpus:
    - feito
    - feyto
    - feitto
    - ffeito
    - ffeyto
  - Regras convertem todas elas para a grafia “feito”
    - Reconhecimento da variação de grafia

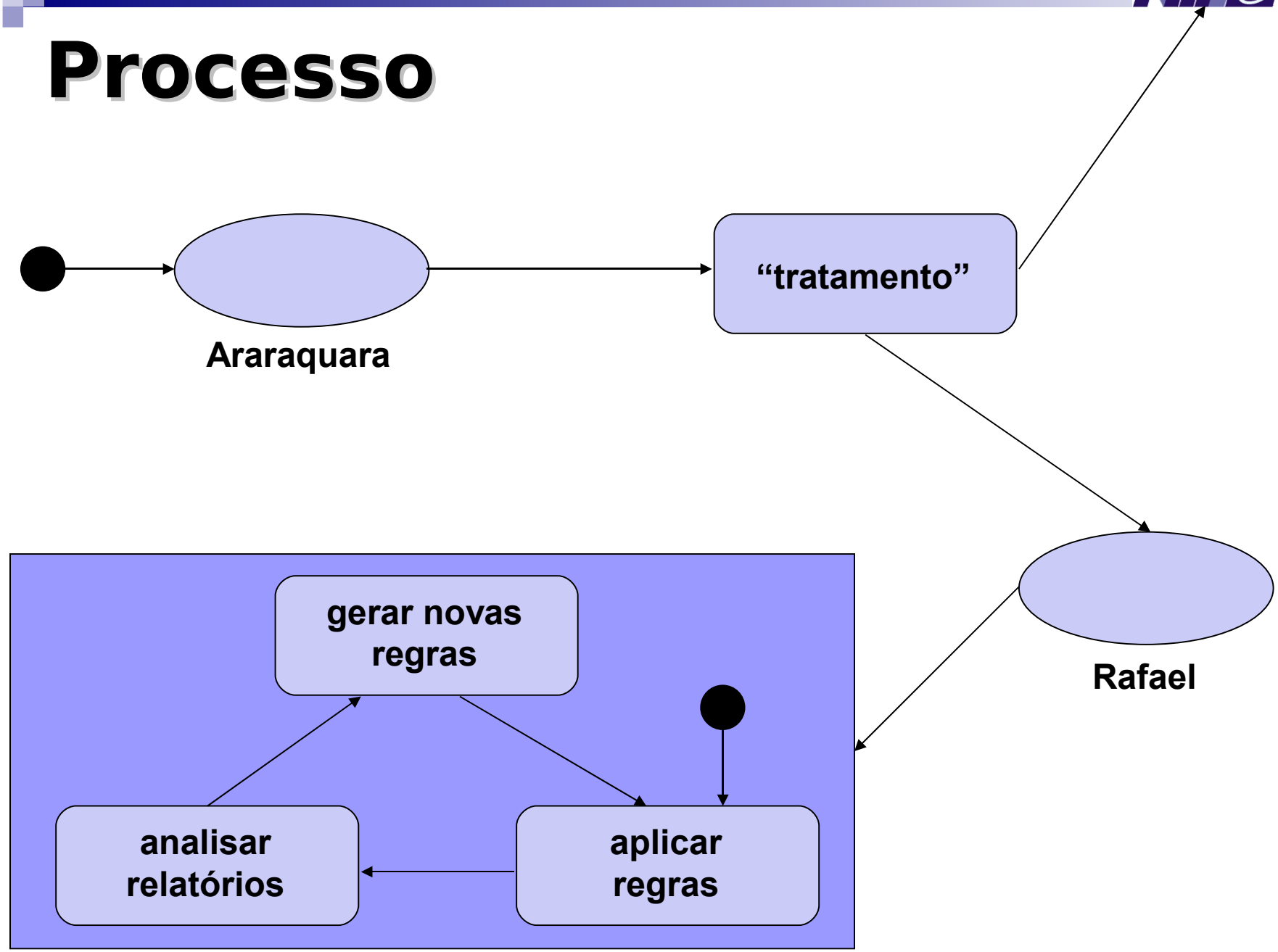
Aplicação

- Nem sempre o agrupamento ocorre em torno de uma ortografia
  - Não estamos preocupados em normalizar, somente detectar variação
  - Podemos eleger uma das grafias como "correta" ou "representativa"
    - Mais freqüente?
    - Escolha arbitrária?
    - Escolha aleatória??
    - Nenhuma???

Aplicação

- Exemplo:
  - As seguintes grafias são encontradas no córpus:
    - tam
    - taõ
    - tao
    - taão
  - Regras convertem-nas para “tam”
    - Agrupamento em torno de uma grafia fora dos padrões ortográficos modernos
      - Mas é um agrupamento

Processo


Araraquara
“tratamento”
Rafael
gerar novas regras
analisar relatórios
aplicar regras

Relatórios

- Relatório de transformações
  - Verificar o efeito sobre cada grafia
- Relatório de agrupamentos
  - Todas as grafias reconhecidas como variação
- Relatório de regras
  - Atuação de cada regra
- Relatório de grafias perdidas
  - Grafias ignoradas pelo conjunto de regras em questão

Relatórios

- Relatório de transformações
  - pedy → pedi
    - (y y i)
  - appellido → apelido
    - (pp pp p)
    - (ll ll l)
  - estaõ → estam
    - (aõ aõ ão)
    - ([^r][aã]o$ [aã]o am)

Relatórios

- Relatório de agrupamentos

```
pedi  (50)
  pedi              (48)
  pedy              (2)

apelido  (48)
  appellido         (23)
  apelido           (19)
  appelido          (6)
```

Relatórios

```
estam  (48)
  estaõ     (29)
  estam     (19)
```

Relatórios

- Relatório de regras
  - (y y i)
    - muyto → muito
    - noyte → noite
    - imdyos → indios
  - (chr chr cr)
    - christo → cristo
    - sepulchro → sepulcro
    - christão → cristao

Relatórios

- Relatório de grafias perdidas

tres	(1.851)
magestade	(1.086)
assucar	(370)
êrros	(1)
çoçobrar	(1)
ácêrca	(1)

Regras em uso

- Eliminação de grafia abandonada
  - y, ph, th...
- Consoantes dobradas
  - ff, pp, dd, gg...
- Normas ortográfias atuais
  - p antes “m” e “b”
  - aã, aõ...
- Regras que visam agrupamento
  - chr → cr christão → cristão
  - .à → á aliàs → aliás

Regras em uso

- Regras lexicalizadas
  - Deos → Deus
  - tres → três
- Regras geradas automaticamente
  - preciz → precis
  - zente → sente
- Todas as regras no RT

Próximos passos

- Seguir analisando os relatórios e gerando regras para agrupamento
- Lexicalizar termos freqüentes
- Melhorar a interface

Fim

- Questionamentos?