Microprojeto

colaboração COGRADE-FILGA

Utilidade para a anotação de topónimos em documentos medievais

REDE TECANDALI ED341D R2016/011 Tecnololoxía e Análise de Datos Lingüísticos

Centro Singular de Investigación en Tecnoloxías da Información

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

- **MOTIVAÇÃO E OBJETIVOS**
- **MATERIAIS E TECNOLOGIAS**
- **RESULTADOS ESPECÍFICOS**
- **TRABALHO FUTURO**

MOTIVAÇÃO E OBJETIVOS

MATERIAIS E TECNOLOGIAS

RESULTADOS ESPECÍFICOS

TRABALHO FUTURO

MOTIVAÇÃO

- Dificuldade na aplicação de ferramentas atuais para anotar textos medievais (trabalho custoso na elaboração manual, porém parcialmente automatizado para textos contemporâneos).
  - Grande variação nos topónimos que limitam a aplicação de gazetteers: *Mondodnedo, Mondonedo, Mondonnedo, Mondonnnedo, Mondoñedo.*
  - Variação que afeta também ao léxico, especialmente relevante na seleção de triggers para o reconhecimento e classificação : *feegresia, figlefia, figressia, figresya, figrigia, figrisia*... até 438 variantes gráficas para um mesmo tipo geográfico!

MOTIVAÇÃO

- Particularidades morfossintáticas que afetam ao PoS tagger e lematização, especialmente relevantes quando são necessárias para as heurísticas de classificação das entidades: Ex. *Na cidade de Santiago* em + artigo + trigger + Nome Próprio -> desambigua em favor da entidade geográfica.

  Nos textos medievais temos também as formas: enno, ẽno, eno, enna, ennas, ...

  Ẽno en+o SPS00+DA0MS0
  porto porto NCMS000
  de de SPS00
  Rriãjo rriãjo NP00G00

MOTIVAÇÃO

- Diversidade de formatos em corpora digitalizados
  - Necessidade de ferramentas específicas para o processamento de metadados, caracteres específicos e modo de anotação para a extração de topónimos.
- Diversidade linguística
  - Num mesmo corpus podem aparecer textos em três línguas e mesmo textos híbridos, de difícil classificação. O corpus tem de ser processado para discriminar por língua em labores específicos de PLN (PoS tagger).

OBJETIVOS

- Mesmo com as dificuldades apontadas, muito do trabalho de anotação de topónimos é mecânico (topónimos que se repetem, estruturas morfossintáticas com alta probabilidade de detetarem a presença de topónimos).
  - É possível adatar uma ferramenta para incrementar a automatização da anotação de topónimos?
  - Qual seria o procedimento mais ajeitado para obter um produto operativo no menor tempo possível?
  - Quais são os elementos determinantes no incremento do desempenho da ferramenta?
  - Qual o incremento obtido face às soluções da língua contemporânea?

MOTIVAÇÃO E OBJETIVOS

MATERIAIS E TECNOLOGIAS

RESULTADOS ESPECÍFICOS

TRABALHO FUTURO

MATERIAIS E TECNOLOGIAS

- Utilidades específicas de córpora
  - Processamento de metadados, caracteres específicos, seleção de língua e modo de anotação para a extração de topónimos.
  - Scripts para criar uma base de dados com metadados e concordâncias (ordenação dos dados).
  - Ambiente web para a visualização dos textos.
  - Base de dados para concordâncias.

MATERAIS E TECNOLOGIAS

Linguakit: Pacote aberto de aplicações PLN com módulo de reconhecimento de entidades por regras.

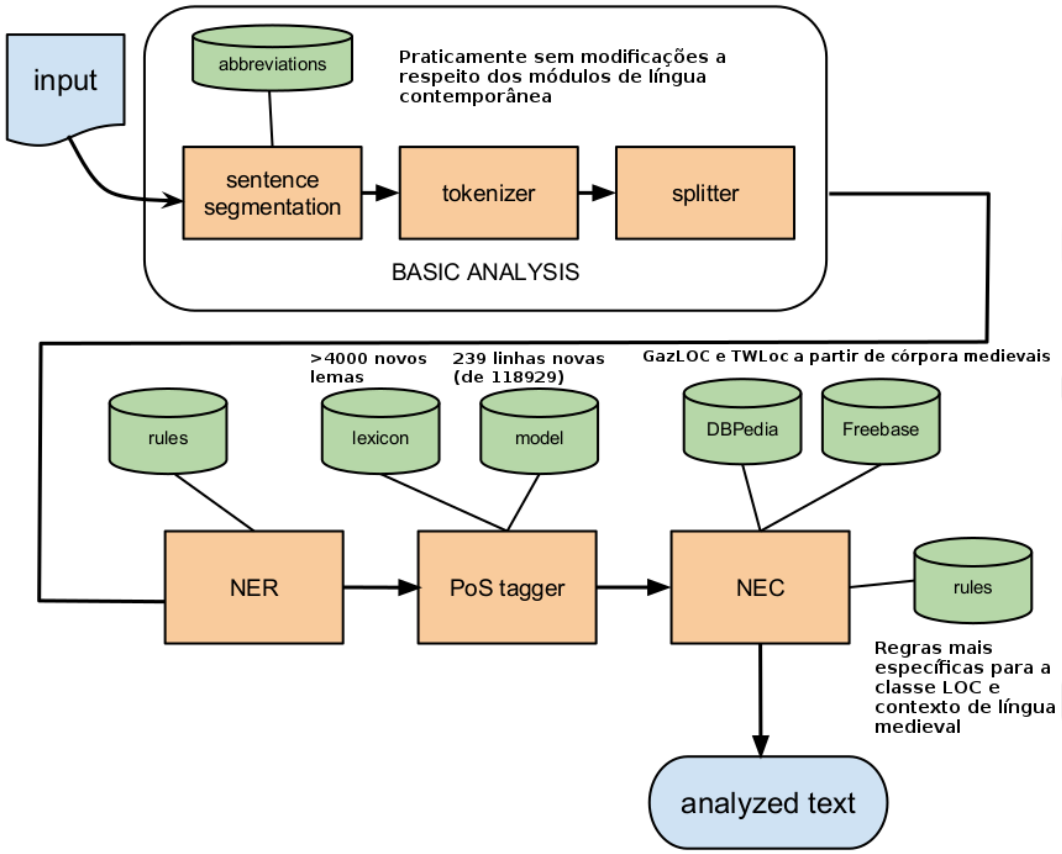


Imagem adatada a partir de Pablo Gamallo (2018)

MATERIAIS E TECNOLOGIAS

- Linguakit

  - Ambiente web para a avaliação de resultados das anotações automáticas. Métricas mais utilizadas: precisão, recall e medida-F.

  - Melhora dos recursos por experimentação (anotação automática, análise de erros, revisão). Prioridades detetadas: gazeteer de topónimos e lista de triggers específicas.

  - Modificação e ampliação de regras mediante a análise de falsos positivos e negativos.

MOTIVAÇÃO E OBJETIVOS

MATERIAIS E TECNOLOGIAS

RESULTADOS ESPECÍFICOS

TRABALHO FUTURO

RESULTADOS ESPECÍFICOS

HISTGZ: utilidadade de reconhecimento de topónimos **completamente** integrado no Linguakit. Disponível em: https://github.com/citiususc/Linguakit

Mais do que anotação de entidades geográficas.

Ẽno en+o SPS00+DA0MS0
porto porto NCMS000
de de SPS00
Rriãjo rriãjo NP00G00
, , Fc
XXVIJ xxvij NP00O00
djas dia NCMP000
de de SPS00
desenbro dezembro NCMS000
, , Fc

RESULTADOS ESPECÍFICOS

HISTGZ: Recursos específicos:

- Lexicon composto da soma dos dicionários de galego e português mais termos com frequência > 20 nos córpora medievais com entidades anotadas (4144 novos lemas).

  sabban saber VMM03P0 saber VMSP3P0

  viren ver VMN03P0 ver VMSF3P0

  vjdas vida NCMP000

  preor prior NCMS000

RESULTADOS ESPECÍFICOS

HISTGZ: Recursos específicos:

- Lista de triggers extraídos do exame do léxico do Corpus Informatizado do Galego-Português Antigo http://ilg.usc.gal/CGPA
  1900 termos geográficos, maiormente variantes.

  ex. *logar, luga, lugar, lugarar, lugare, lughar, luguar , luguoar, luguuares...*

RESULTADOS ESPECÍFICOS

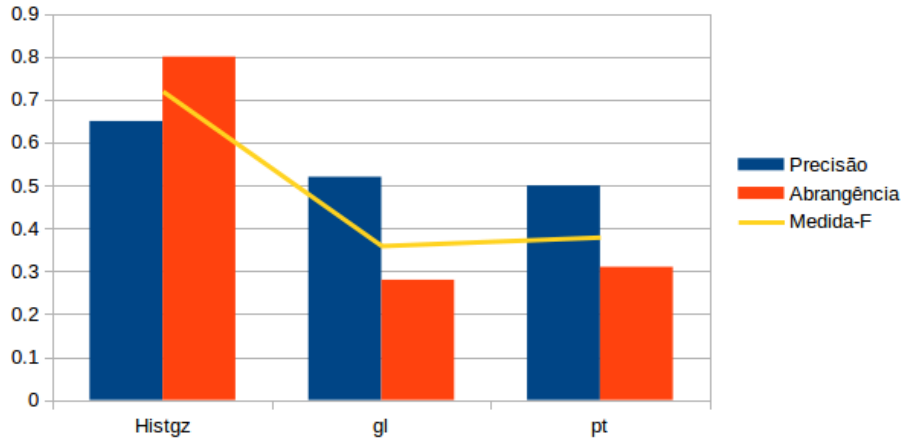
HISTGZ: Recursos específicos adicionados ao módulo de língua contemporânea:

- Gazetteer de topónimos medievais extraída dos córpora anotados (5989 topónimos).
- Modelo do PoS tagger a partir do galego, expandido com treino num corpus reduzido (2022 linhas) representativo das distintas variedades dos córpora anotados.
- Regras específicas para a melhora da anotação de topónimos e língua medieval.

RESULTADOS ESPECÍFICOS

Comparativa de desempenho dos módulos HISTGZ, Galego e Portugués no Linguakit

Características do texto a anotar: extraído aleatoriamente sobre corpus não usado para gazetteers nem testes de avaliação. 2060 tokens com 35 topónimos validados manualmente.



Desempenho de módulos do Linguakit na anotação de entidades geográficas mencionadas num texto galego medieval

RESULTADOS ESPECÍFICOS

Conclusões de melhora

Por especificações na heurística

morador morador NCMS000
ẽna en+a SPS00+DA
freigresja fregresia NCFS000
de de SPS00
Santa_Coõba_de_Rriãjo santa_coõba_de_rriãjo NP00SP0

Regras mais específicas para a desambiguação de entidades.

Por abrangência da gazetteer

cárregos cárregos AQ0000
rrey rei NCMS000
e e CC
Rroma rroma NP00G00
e e CC
See see NP00SP0

Consideração da variação gráfica na gazetteer.

RESULTADOS ESPECÍFICOS

Conclusões de melhora

Por limitações na lematização e abrangência de listas complementares

a o DA0FS0
dorna dorna NCFS000
de de SPS00
XXVJ xxvj NP00G00
canadas canada NCFP000
vellas vello AQ0FP0

Gonçaluo_Mariño gonçaluo_mariño NP00V00
e e CC
Rroý_Mariño rroý_mariño NP00V00
, , Fc
escudeyros escudeyros AQ0000
de de SPS00
Sueyro_Gomes sueyro_gomes NP00G00

Aumentar o lexicon com lemas mais frequentes e sistematizáveis (ex. numerais romanos, classificados como nomes próprios por começarem com maiúscula).

Gazetteer específica de nomes de pessoa medievais.

RESULTADOS ESPECÍFICOS

Conclusões de melhora

Pelo critério utilizado para a classificação da entidade

Testigos testigos NP00V00
: : Fd
Vasco vasco AQ0MS0
de de SPS00
Lees lees NP00G00
e e CC
Martjn_de_Tourís martjn_de_tourís NP00V00

Johán_Peres johán_peres NP00SP0
de de SPS00
Põtevedra põtevedra NP00G00
Pero_Santiagujño pero_santiagujño NP00SP0

Pero_de_Villanustre pero_de_villanustre NP00SP0
por por SPS00
si si PP3CSO00
e e CC
en en SPS00
nome nome NCMS000
de de SPS00
Johán_Pereyreyros johán_pereyreyros NP00SP0

Necessidade de homogeneizar os critérios de anotação mais linguísticos com os padrões NERC. Qual é a entidade referente?

MOTIVAÇÃO E OBJETIVOS

MATERIAIS E TECNOLOGIAS

RESULTADOS ESPECÍFICOS

TRABALHO FUTURO

RESULTADOS ESPECÍFICOS

Conclusões de melhora

Pelo critério utilizado para a classificação da entidade

Testigos testigos NP00V00
: : Fd
Vasco vasco AQ0MS0
de de SPS00
Lees lees NP00G00
e e CC
Martjn_de_Tourís martjn_de_tourís NP00V00

Johán_Peres johán_peres NP00SP0
de de SPS00
Põtevedra põtevedra NP00G00
Pero_Santiagujño pero_santiagujño NP00SP0

Pero_de_Villanustre pero_de_villanustre NP00SP0
por por SPS00
si si PP3CSO00
e e CC
en en SPS00
nome nome NCMS000
de de SPS00
Johán_Pereyreyros johán_pereyreyros NP00SP0

Necessidade de homogeneizar os critérios de anotação mais linguísticos com os padrões NERC. Qual é a entidade referente?

TRABALHO FUTURO

Específico do HISTGZ

- Introducir triggers e gazetteers para as entidades não geográficas.
- Melhorar a base de treino do modelo validando corpora processados automaticamente pelo Linguakit.
- Utilizar a análise de erros das validações para melhorar as heurísticas de classificação de entidades.
- Ampliar o lexicon e corrigir e melhorar a lematização.

Da anotação de topónimos

- Comparativa com sistema estatístico (possível uso das concordâncias para treinar um sistema de ML).
- Base de dados para georreferenciação e desambiguação geo / geo.

Obrigado

Participaram no desenvolvimento: José Angel Taboada, Xavier Varela, Xosé Ramón Viqueira, Pablo Gamallo, David Mera, Paulo Martínez Lema, Xavier Canosa

REDE TECANDALI ED341D R2016/011 Tecnololoxía e Análise de Datos Lingüísticos

Grupos
FILGA (Filoloxía e Lingüística Galega GI-1743)
COGRADE (Gráficos por Computador e Enxeñaría de Datos GI-2116).

galicia

UNIÓN EUROPEA

XUNTA DE GALICIA

Fondo Europeo de Desenvolvemento Rexional *"Unha maneira de facer Europa"*