Processamento Superficial

Parte 2/3

trabalho conjunto com
João Silva
Filipe Nunes
Pedro Martins

Tecnologia da linguagem

- Aplicações específicas
  - correcção ortográfica e gramatical
  - sumarização
  - aprendizagem de línguas assistida por computador
  - resposta-a-perguntas
  - interfaces em língua natural
  - tradução assistida por computador
  - tradução automática
  - ...
- Integração e suporte a outras aplicações
  - acessibilidade por deficientes
  - recuperação de texto
  - prospecção de informação
  - eLearning
  - localização e multilinguismo
  - ...
- Web semântica
  - Anotação automática de meta-dados

Da forma ao significado

- Objectivo
  - Explicitar e associar informação linguística ao texto
- Cadeia de processamento
  - reconhecimento da fala
  - segmentação de lexemas e frases
  - anotação morfo-sintáctica
  - análise morfológica nominal
  - análise morfológica verbal
  - lematização
  - reconhecimento de entidades nomeadas
  - reconhecimento de expressões multi-palavra
  - desambiguação de acepções
  - análise sintáctica (constituência)
  - funções gramaticais
  - papéis semânticos
  - forma lógica
  - ...
- Problemas chave: ambiguidade e novidade

Segmentação de frases

- **Dificuldades chave**
  - Ambiguidade e ambivalência do ponto: `Sr. João`
  - Ambiguidade do travessão: `— Eu — sussurrou — também.`
- **Abordagem**
  - Expressões regulares, e.g. "ponto seguido de maiúscula"
  - Lista de abreviaturas
- **Contribuição**
  - Cobertura exaustiva, para além de *proof-of-concept*
- **Avaliação (Medida-F)**
  - Valor base: 99,31% Ferramenta: **99,94%**
    - 12 000 frases (incl. 18% diálogo)
    - Valor base: segmenta sempre que "`. X`"

Segmentação de lexemas

- **Dificuldades chave**
  - Ambiguidade de sequências: `deste, consigo, pelas...`
- **Abordagem**
  - Segmentação e anotação interpoladas
    - Manter sequências ambíguas: `deste → |deste|`
    - Anotação em contexto: `|deste/PREPDEM|` ou `|deste/V|`
    - Expansão: `|deste/PREPDEM| → |de/PREP|este/DEM|`
- **Contribuição**
  - Primeira ferramenta a resolver sequências em contexto
- **Avaliação**
  - Valor base: 98,46% Ferramenta: **99,72%**
    - Medição apenas sobre sequências ambíguas: 5450 tokens
    - Valor base: leitura mais provável

Anotação morfo-sintáctica

- **Dificuldades chave**
  - Ambiguidade categorial: `como ← ADV|CJ|INT|PREP|REL|V`
  - Novidade: `concordanciador, ...`
- **Abordagem**
  - Métodos estado-da-arte: Aprend. p/ tranformações (TBL, Brill), Máxima entropia (MXPOST, Ratnaparkhi), Modelos de Markov (TnT, Brants)
- **Contribuição**
  - Desempenho ao nível do estado-da-arte para outras línguas
  - Melhores anotadores para o português (Branco e Silva, 2004)
- **Avaliação (correcção)**
  - Valor base (TnT): 96,37% Ferramenta (TnT): **98,52%**
    - Avaliação sobre 260K, 10% não usados para treino (x10 iterações)
    - Valor base: categoria mais provável

Traçamento nominal

- **Dificuldades chave**
  - Ambiguidade de traços: `pianista` (f/m?), `lápis` (s/p?)...
  - Novidade: `traflemadora` (f/m?; s/p?), ...
- **Abordagem**
  - Subst de terminações x lista de excepções
  - Explorar concordância: `o lápis` vs. `os lápis`
- **Contribuição**
  - Primeira ferramenta dedicada
  - Baseado em regras com léxico mínimo
  - Resolução da ambiguidade em contexto
- **Avaliação**
  - Valor base: 87,08% Ferramenta: **92,62%**
  - Com processamento sintáctico subsequente correcto: 98,79%
    - Medição sobre anotação morfo-sintáctica correcta
    - Valor base: sem propagação de concordâncias

Lematização nominal

- **Dificuldades chave**
  - Ambiguidade semântica: `ética` (ética/o?), `copas` (copa/s?)
  - Novidade: `traflemadoras` (-dor,-dora,...?)...
- **Abordagem**
  - Subst de terminações x Lista de excepções
  - Minimização da lista de excepções
- **Contribuição**
  - Primeira ferramenta dedicada
  - Baseado em regras com léxico mínimo
- **Avaliação**
  - Valor base: 87,08% Ferramenta: **97,67%**
    - Avaliação sobre 51K nominais
    - Melhoria expectável: interpolação com WSD
    - Valor base: só com regras de terminação

Lematização e traçamento verbais

- **Dificuldades chave**
  - Ambiguidade de traços e/ou lema:
    - `virei(vir/Fut1s; virar/Perf1s), parti(partir/Pres1s/Imp2p)`
  - Novidade: `concordancear`
- **Abordagem**
  - Motor de flexão: subst de terminações x lista de excepções
  - Máxima verosimilhança em cascata
- **Contribuição**
  - Primeira ferramenta dedicada
  - Resolução da ambiguidade
- **Avaliação**
  - Valor base: 88,52% Ferramenta: **95,92%**
    - Sobre 40K verbais e anotação morfo-sintáctica correcta
    - Valor base: atribuir o conjunto de traços mais provável
    - Trabalho em curso: classificador Bayesiano e SVM

Serviços online

- Já disponíveis
  - Suite
    - `lxsuite.di.fc.ul.pt`
  - Flexão nominal
    - `lxinflector.di.fc.ul.pt`
  - Flexão verbal
    - `lxlemmatizer.di.fc.ul.pt`
    - `lxconjugator.di.fc.ul.pt`
- Serviços online em finalização
  - Concordanciador
  - Reconhecedor de entidades nomeadas