Anotação de Corpus

Sandra Maria Aluisio

Primórdios

- Doutorado, 1991: **Anotação retórica** de artigos científicos (54 introduções) no projeto AMADEUS
  - Suporte a criação de ferramentas de escrita
  - ✓ **Adhoc**: uso do modelo CARS (Swales, 1990) e não um manual, com anotação feita por 1 única pessoa, sem chances de calcular a concordância da anotação, MAS com um padrão ótimo de intercâmbio (SGML).

(1991-1995): AMADEUS

2003: SciPo

2004: SciPo-Farmácia

2011: MAZEA-WEB Anotação multirrótulo

Aluísio, S.M. and Oliveira, Jr. O.N. A Case-Based Approach for Developing Writing Tools Aimed at Non-native English Users. Lecture Notes In Computer Science 1010, pp. 121-132 (ICCBR' 95), 1995.

Primórdios

- NILC desde sua criação em 1993: **Anotação de erros gramaticais** (desvios com relação à norma culta)
  - corpus (11.624 sentenças, 2616 com erros) contendo sentenças e marcas de erro e de tipo de erro - o Probi – para evitar falsos positivos, principalmente, no revisor gramatical ReGra, de forma automática.
  - Suporte a criação do corretor gramatical ReGra
  - Usado para avaliar outro corretor, o CooGroo (http://ccsl.ime.usp.br/cogroo/maven/resultados/resumos/FMeasure-PROBI.html)

Martins, R.T. (2002) *PROBI: um corpus de teste para o revisor gramatical ReGra*. NILC-TR-02-10, 7p.

Definição

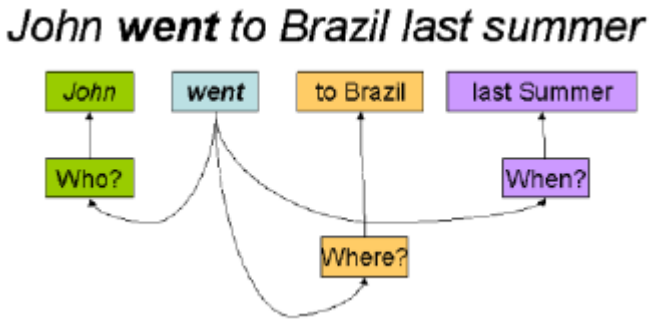
- Anotação (‘tagging’) é o processo de adicionar novas informações em textos fontes, seja por humanos (anotadores) ou por sistemas treinados para a tarefa (anotação automática)

- Decisão:
  - Material a ser anotado
  - Teoria/conhecimento que o anotador possui, seja porque foi treinado para isto ou adquiriu previamente

Processo de Anotação

- Decidir que fragmento do texto anotar
- Adicionar uma etiqueta, de um conjunto fixo, pré-definido (tagset)

Tagging Semântico/ Decisão



Tagging Morfossintático/ Tagset

Pesquisadores_N
de_PREP|+
a_ART
Aids_N
em_PREP|+
os_ART
EUA_NPROP
identificaram_V
a_ART
substância_N

ADJETIVO	ADJ
ADVÉRBIO	ADV
ADVÉRBIO CONECTIVO SUBORDINATIVO	ADV-KS
ADVÉRBIO RELATIVO SUBORDINATIVO	ADV-KS-REL
ARTIGO (def. ou indef.)	ART
CONJUNÇÃO COORDENATIVA	KC
CONJUNÇÃO SUBORDINATIVA	KS
INTERJEIÇÃO	IN
NOME	N
NOME PRÓPRIO	NPROP
NUMERAL	NUM
PARTICÍPIO	PCP
PALAVRA DENOTATIVA	PDEN
PREPOSIÇÃO	PREP
PRONOME ADJETIVO	PROADJ
PRONOME CONECTIVO SUBORDINATIVO	PRO-KS
PRONOME PESSOAL	PROPESS
PRONOME RELATIVO CONECTIVO SUBORDINATIVO	PRO-KS-REL
PRONOME SUBSTANTIVO	PROSUB
VERBO	V
VERBO AUXILIAR	VAUX
SÍMBOLO DE MOEDA CORRENTE	CUR
CONTRAÇÕES e ÊNCLISES	|+
MESÓCLISES	|!
ETIQUETAS COMPLEMENTARES	
Estrangeirismos	|EST
Apostos	|AP
Dados	|DAD
Números de Telefone	|TEL
Datas	|DAT
Horas	|HOR
Disjunção	|[ ]

fsp10.s4: [A neblina] **atrapalhou** outra_vez [as operações de o aeroporto] [por mais_de oito horas] .

O quê atrapalhou? [A neblina]
Atrapalhou o quê? [as operações de o aeroporto]
Atrapalhou por quanto tempo? [por mais_de oito horas]

????? [**outra vez**]

Tipos de anotação

- In-line: todas as anotações estão no mesmo arquivo fonte
- ✓ **Standoff**: cada tipo/nível de anotação em arquivos separados e o arquivo fonte não possui anotação
  - Permite trocar a anotação de um nível (p.ex. trocar de etiquetador) sem alterar as outras anotações.
    - Usada no projeto **PLN-BR do NILC** (PLN-BR GOLD, CATEG e FULL): (http://www.nilc.icmc.usp.br:8180/portal/)
    - Córpus SUMM-IT com anotação anotação morfossintática (automática), anotação de co-referência dos sintagmas nominais (manual) e anotação de relações retóricas, ou relações RST (manual).

Porquê anotar?

- **PLN:**
  - Alguns fenômenos linguísticos/tarefas são muito complexas para serem definidas usando regras
  - Para dar conta desta complexidade, na área de PLN se aplicam métodos que aprendem a partir de corpora anotados – APRENDIZADO DE MÁQUINA
    - Anotação serve com insumo para alimentar os métodos de aprendizado
  - Metodologia:
    - Vários anotadores humanos anotam um corpus
    - Avaliação da concordância da anotação usando estatística como o kappa
    - Uso de um método de aprendizado de máquina para a tarefa

Porquê anotar?

- **Linguistas:**
  - Permite a busca por fenômenos linguísticos
  - Gerar estatísticas para o fenômeno
  - Descobrir novos fenômenos e correlações
  - Testar uma teoria

Quais fenômenos anotar para a LP?

- Advérbios e locuções adverbiais de tempo, lugar, quantidade, causa, que inclusive melhorariam parsers
- Opiniões (polaridade) e sentimentos (esta área está em foco com o Twitter e a Pesquisa de Opinião na Web)
- Complexidade de textos de acordo com séries/anos escolares ou níves do INAF, para fundamentar políticas nacionais
- E muito mais...

Importância de Corpora anotados

- **Corpora são mais importantes do que métodos computacionais**
  - Bons corpora anotados duram décadas; métodos são substituidos por novos métodos, mais rapidamente
    - Penn Treebank Project (1989-1992)
    - Mac Morpho do Projeto Lácio-Web (2004)
  - Um projeto de corpus mal conduzido, pode prejudicar a pesquisa de uma área por anos

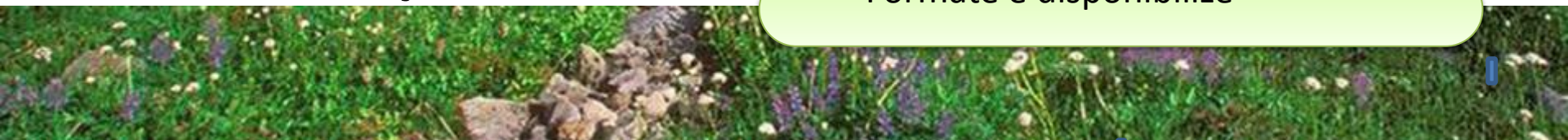
O que se espera da anotação

- Anotação deve ser:
  - **Rápida** … para produzir um grande corpus
  - **Consistente**… para permitir aprendizado de máquina
  - **Profunda**… o bastante para ser interessante

- É necessário:
  - Uma **metodologia** simples e uma boa **interface** de anotação
  - **Várias pessoas anotando** para não permitir tendências de um único anotador
  - Atenção com a **teoria** que está por traz da anotação
  - Uso de um bom **padrão** para intercâmbio de dados
  - **Distribuição** eficiente destes recursos caros para que sejam **reusáveis**
    - muitas vezes só servem para uma aplicação, um grupo de pesquisa

Estágios de um Projeto de Anotação

- Seleção da Tarefa
  - Escolha um problema
  - Tome decisões iniciais
  - Produza um manual
  - Teste a tarefa com pessoas
- Preparação
  - Colete o corpus
  - Escolha ou construa uma interface para anotação
  - Contrate anotadores e gerentes
- Anotação
  - Treine anotadores
  - Faça a anotação
  - Monitore progresso e concordância
  - Faça encontros periódicos
- Avaliação
  - Avalie desempenho
  - Rode testes com métodos de aprendizado de máquina
- Distribuição
  - Formate e disponibilize



Agenda

- Projetos de anotação para o PB desenvolvidos pelo NILC

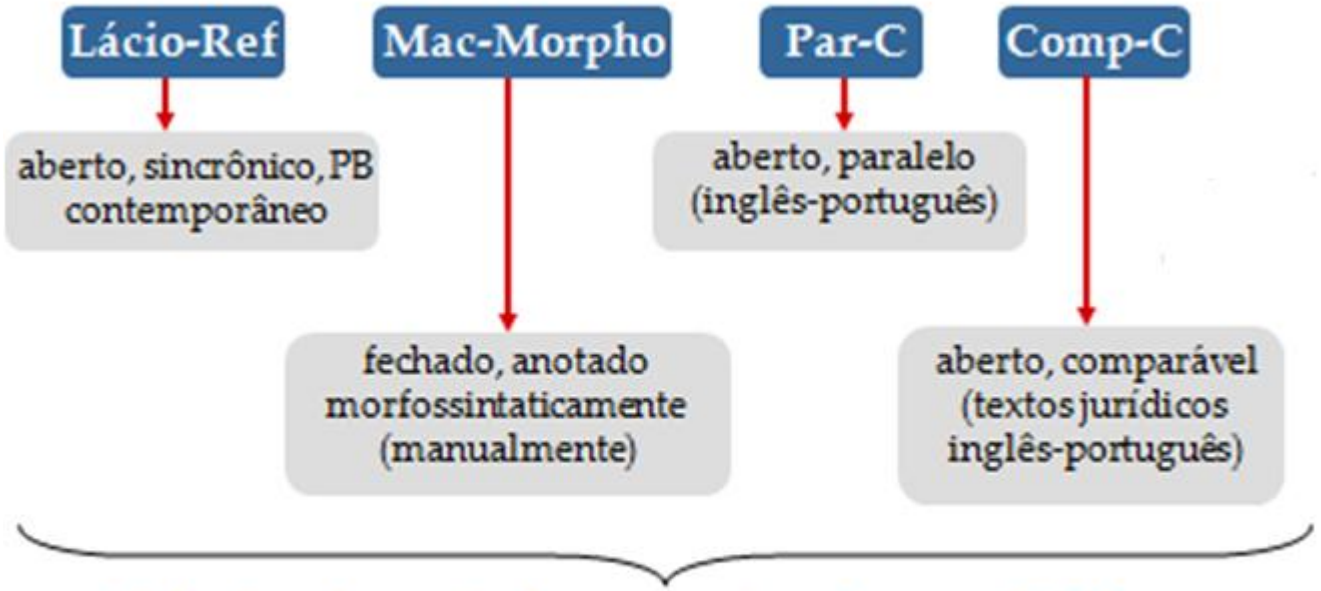
- Questões em aberto desta área
  (1) Qual Corpus? Como conseguir um corpus balanceado para anotar? Quando o corpus é balanceado, representativo e ainda atual (não defasado)?
  (2) Como permanecer fiel à teoria? Como escrever um bom manual (não é trivial).
  (3) Quais as características importantes na seleção de anotadores? Como garantir que estão treinados de forma adequada?
  (4) Como criar um procedimento de anotação simples, rápido e confiável?
  (5) Que interfaces são melhores para cada tipo de problema e como garantir que elas não influenciam os resultados?
  (6) Como avaliar os resultados da anotação? Quais medidas de concordância são apropriadas?
  (7) Como armazenar os resultados?; Quando e para quem disponibilizar o corpus? Questões de licença, manutenção e distribuição.

Agenda

- **Projetos para o PB criados pelo NILC**

- Questões em aberto desta área
  (1) Qual Corpus? Como conseguir um corpus balanceado para anotar? Quando o corpus é balanceado, representativo e ainda atual (não defasado)?
  (2) Como permanecer fiel à teoria? Como escrever um bom manual (não é trivial).
  (3) Quais as características importantes na seleção de anotadores? Como garantir que estão treinados de forma adequada?
  (4) Como criar um procedimento de anotação simples, rápido e confiável?
  (5) Que interfaces são melhores para cada tipo de problema e como garantir que elas não influenciam os resultados?
  (6) Como avaliar os resultados da anotação? Quais medidas de concordância são apropriadas?
  (7) Como armazenar os resultados?; Quando e para quem disponibilizar o corpus? Questões de licença, manutenção e distribuição.

Anotação Morfossintática no Lácio-Web



<ppar=966418>
<s>
Para_PREP
se_PROPESS
instalar_V
em_PREP|+
a_ART
Vila=Olímpia_NPROP
contratou_V
os_ART
serviços_N
de_PREP|+
o_ART
elegante_ADJ
arquiteto_N
Aurelio=Martinez
=Flores_NPROP
._.
</s>
</p>

Anotação morfossintática no Lácio-Web

- **Seleção da Tarefa**:
  - Motivação: Falta de um grande corpus para treinar classificadores
  - Qual tagset usar? Eagles recommendations for the Morphosyntactic Annotation of Corpora (http://www.ilc.pi.cnr.it)
- **Preparação**:
  - Corpus de textos jornalísticos (Variedade de Tópicos Cadernos)
  - REVISAR uma anotação automática é mais rápida do que anotar em corpus crú
  - Anotação em TXT, uma palavra por linha, sem editor de anotação
  - Anotadores: 4 linguistas, um deles senior (gerente)
  - Criação de um MANUAL, com vários EXEMPLOS POSITIVOS e NEGATIVOS

- **Anotação:**
  - Reuniões SEMANAIS para discutir dúvidas
  - Revisão do Manual: 10 versões
  - Revisão dos textos para cada mudança

- **Avaliação:**
  - Avaliação da concordância com a estatística KAPPA (0.944 e 0.955)
  - Tempo de anotação: 11 meses
  - Criação de 3 taggers no projeto (http://www.nilc.icmc.usp.br/lacioweb/ferramentas.htm)

- **Distribuição:**
  - Formato para pesquisas linguísticas
  - Formato para treinamento de taggers: 1.2 milhões de palavras
  - Portal do Lácio-Web: http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm
  - Usado como benchmark para a tarefa
    - p. ex. usado para avaliar o método Entropy Guided Transformation Learning, publicado no PROPOR 2008

ALUÍSIO, S. M.; PELIZZONI, J. M.; MARCHI, A. R.; OLIVEIRA, L. H.; MANENTI, R.; MARQUIVAFÁVEL, V. (2003). An account of the challenge of tagging a reference corpus of Brazilian Portuguese. In: PROPOR´2003, 2003, Faro. Lecture Notes on Artificial Intelligence. Proceedings of PROPOR´2003. Springer Verlag, 2003. v. 1.

::: MAC-Morpho - Concordanciador para córpus anotado morfossintaticamente

O concordanciador implementado gera uma lista enumerada de todas as ocorrências
conjunto de etiquetas do Mac-Morpho.
São também opções de escolha:
- o córpus onde tal expressão ou palavra será procurada: MAC-MORPHO em sua tota
- o tamanho, em caracteres, do contexto reduzido (esquerdo e direito), na lista de oco
- o tamanho, em caracteres, do contexto expandido, onde a expressão ou palavra occ
Essa última opção dita o contexto, da expressão ou palavra, que é visualizado após o
em torno das ocorrências da expressão ou palavra escolhida.

Expressão ou Palavra: Igual a

Etiqueta: Todas

Escolha o Córpus: MAC-MORPHO

Opções:

Diferenciar maiúsculas e minúsculas: ☑

Tamanho do contexto reduzido: 20

Tamanho do contexto expandido: 100

Executar Limpar

Foram encontradas 6386 ocorrências!!
Fazer DOWNLOAD do arquivo resultado do concordanciador.

```
1  ,_, que_PRO-KS-REL se_PROPESS aposentou_V
2 igor_N ,_, nota_V|+ se_PROPESS um_ART aper
3  ,_, mas_KC não_ADV se_PROPESS realizava_V
4 P|+ os_ART locais_N se_PROPESS manteve_V c
5 ADJ ,_, deve_VAUX|+ se_PROPESS aguardar_V
6 igar_V somente_PDEN se_KS o_ART total_N de
7  ciclo_N tornar_V|+ se_PROPESS longo_ADJ .
8 EP colheitadeiras_N se_PROPESS concentram_
9  safra_N 93/94_N|AP se_PROPESS inscreveram
10 e_KC que_PRO-KS-REL se_PROPESS inscreveram
11 diz_V que_KS já_ADV se_PROPESS pode_VAUX p
12 ra_PREP pastagens_N se_PROPESS caracterizo
13 e_KC que_PRO-KS-REL se_PROPESS adapta_V be
14  não_ADV podia_VAUX se_PROPESS dar_V bem_A
15  terra_N tornou_V|+ se_PROPESS praticament
16 _NPROP ._. Joga_V|+ se_PROPESS muita_PROAD
17 rção_N e_KC não_ADV se_PROPESS estudam_V o
18  ,_, costuma_VAUX|+ se_PROPESS colocar_V s
19 _ADJ que_PRO-KS-REL se_PROPESS tornarão_V
20 ão_V desérticos_ADJ se_KS não_ADV tiverem_
21 ) ) , , acertou_V|+ se_PROPESS que_KS os_A
```

Anotação de Papéis Semânticos no Propbank-Br

- Anotação de Papéis Semânticos:

  1) identificação do "argument taker"/evocador, que pode ser um único verbo ou um predicado complexo (light verb constructions ou phrasal verbs, p.ex.);

  2) identificação e delimitação de argumentos associados com o evocador, e

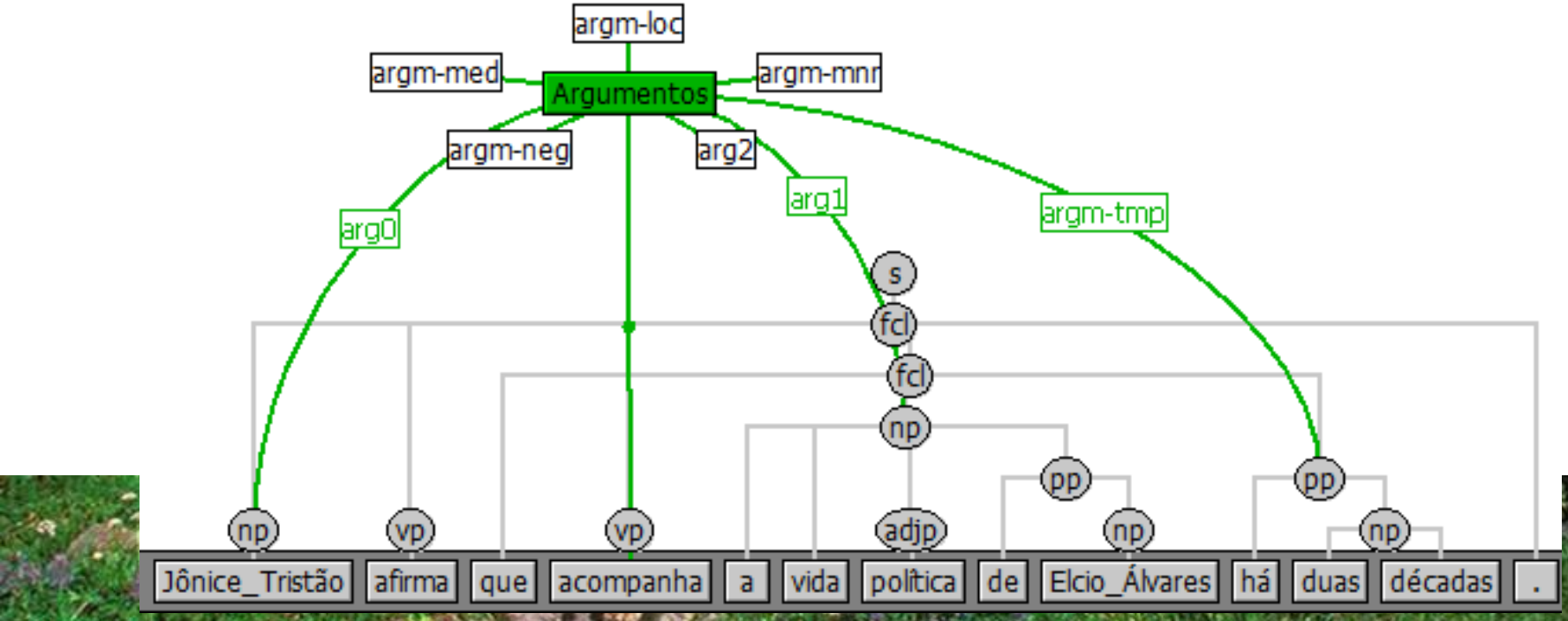
  3) atribuição de um papel semântico para cada um destes argumentos.

- Duplicação das sentenças para anotar cada verbo:

- SENTENÇA 1: O aumento de casos fez reverem estratégias.
- SENTENÇA 1 A (Fazer). O aumento de casos **fez** reverem estratégias.
- SENTENÇA 1.B (Rever). O aumento de casos fez **reverem** estratégias.

- Anotação em cima da árvore sintática elimina o passo de delimitação de argumentos.

- Porém, a qualidade da anotação é dependente da qualidade do parser



- **Seleção da Tarefa**:
  - Motivação: Falta de um corpus para treinar taggers semânticos
  - Qual teoria usar? Propbank: que usa uma teoria neutra para papéis semânticos (ArgsN e ArgsM) que se mostrou eficaz para o treinamento
- **Preparação**:
  - Corpus anotado sintaticamente, com precisão. Escolha do corpus BOSQUE, que foi revisado. Problema: tamanho e sentenças AINDA com erro de anotação. Porção Brasileira possui 4213 sentenças. Folha de São Paulo do ano de 1994.
  - Avaliação de várias ferramentas públicas para anotação: escolha do SALTO (http://www.coli.uni-saarland.de/projects/salsa/salto/doc/)
  - Adaptação do manual do Propbank para a língua portuguesa
  - Duplicação das sentenças para anotar um verbo por vez (automatizado): 7107 instâncias de anotação e 1068 diferentes verbos plenos
  - Excluídos os verbos auxiliares com base em uma tabela de verbos auxiliares

- **Anotação:**
  - Anotadores: **1 único** (Projeto Pós-doc de MAGALI SANCHES)
  - Para lidar com os desafios da LP, incrementou-se a anotação com "sentence flags" para marcar todas as ocorrências de:
    - orações reduzidas / partícula "se" pronominal / sujeito oculto
    - sujeito indeterminado / elipse / correferência
    - predicados complexos / multipalavras não reconhecidas pelo parser

- **Avaliação:**
  - Não pode ser feita a avaliação da concordância, pois só havia um anotador
  - Tempo de anotação: 9 meses
  - Será feita via criação de taggers em outros projetos do NILC
  - 6142 instâncias anotadas e 1068 predicados verbais diferentes

- **Distribuição:**
  - PortLex (http://www2.nilc.icmc.usp.br/portlex/)

DURAN, M. S.; ALUÍSIO, S. M. (2011) Propbank-Br: a Brazilian Portuguese corpus annotated with semantic role labels. In the Proceedings of The 8th Brazilian Symposium in Information and Human Language Technology (STIL 2011), Cuiabá-MT, CD-ROM, v. 1, ISSN 2175-6201, pp. 164-168.

Outras tarefas

- Anotação de **Operações de Simplificação Sintática e Léxica** no Projeto PorSimples (**http://caravelas.icmc.usp.br/wiki/**) (2007-2010) em textos jornalísticos

  – Simplificação natural e forte, usado para aprender a tarefa de simplificação

  – Criação de uma interface para a tarefa

  – **PROBLEMA**: anotação feita por um único anotador

Outras tarefas

- Anotação da **Estrutura Retórica de dois grandes corpora de resumos de artigos científicos (2010-2011) (http://www.nilc.icmc.usp.br/mazea-web/)** com multirrótulos para cada oração – tarefa nova para este nível de anotação que não é sentencial

  – Uso de um etiquetador monorrótulo, cuja anotação foi revisada

  – **SOLUÇÃO do Problema de uso de 1 único anotador:** avaliação da anotação feita por vários anotadores, via estatística KAPPA para refinar o manual (apontar pontos de discordância) e posterior anotação do corpus feita por 1 único anotador

Agenda

- Projetos para o PB criados pelo NILC

- Questões em aberto desta área
  - **(1) Qual Corpus? Como conseguir um corpus balanceado para anotar? Quando o corpus é balanceado, representativo e ainda atual (não defasado)?**
  - (2) Como permanecer fiel à teoria? Como escrever um bom manual (não é trivial).
  - (3) Que interfaces são melhores para cada tipo de problema e como garantir que elas não influenciam os resultados?
  - (4) Quais as características importantes na seleção de anotadores? Como garantir que estão treinados de forma adequada?
  - (5) Como criar um procedimento de anotação simples, rápido e confiável?
  - (6) Como avaliar os resultados da anotação? Quais medidas de concordância são apropriadas?
  - (7) Como armazenar os resultados?; Quando e para quem disponibilizar o corpus? Questões de licença, manutenção e distribuição.

Q1: Preparação – Escolha do Corpus

- Escolha deve ser feita com cuidado, pois espera-se que o corpus seja muito reusado
  - Durar 30 anos!
  - Vejam o caso do Penn Treebank que usou uma seção de finanças.

- Balanceamento: gênero/era/domínio
  - Explique a razão da escolha do balanceamento

- Uma solução: começar com o que se tem disponível e balancear numa segunda etapa.
  - Como lidar com os novos gêneros de textos que surgiram com a Web, quais são eles??

- Não temos no Brasil um distribuidor/ concentrador de corpus como o LDC (Linguistic Data Consortium) — www.ldc.upenn.edu

  - **Não seria o caso de começar este consórcio?**

Agenda

- Projetos para o PB criados pelo NILC

- Questões em aberto desta área
  (1) Qual Corpus? Como conseguir um corpus balanceado para anotar? Quando o corpus é balanceado, representativo e ainda atual (não defasado)?
  (2) Como permanecer fiel à teoria? Como escrever um bom manual (não é trivial).
  (3) Que interfaces são melhores para cada tipo de problema e como garantir que elas não influenciam os resultados?
  (4) Quais as características importantes na seleção de anotadores? Como garantir que estão treinados de forma adequada?
  (5) Como criar um procedimento de anotação simples, rápido e confiável?
  (6) Como avaliar os resultados da anotação? Quais medidas de concordância são apropriadas?
  (7) Como armazenar os resultados?; Quando e para quem disponibilizar o corpus? Questões de licença, manutenção e distribuição.

Q2: Instanciação da teoria

- Detalhe da anotação (sofisticação) versus Sucesso da Anotação
  - Faça testes para determinar o que será anotado na prática
  - Antes de anotar não dá para saber quão fácil os anotadores vão identificar as categorias da teoria

- Especialistas criam um manual e dizem quais as categorias

- MAS não devem congelar o manual muito cedo
  - Especialistas anotam uma amostra e medem concordância – *gold standard*
  - Anotadores anotam a amostra até que o esquema seja compreendido

- Faça reuniões semanais com anotadores e meça a concordância da anotação
  - Atualize o manual com os casos especiais

Precisão versus Kappa

- Se houver um **gold standard** avaliar a **precisão** da anotação, que é preferível do que **kappa**
  - Precisão diz qual a facilidade de anotar as categorias decididas
  - Crie 2 classes (classe de interesse e todas as outras colapsadas) e calcule a concordância
  - Repita para todas as classes

- Se a teoria estiver emperrando a anotação
  - Neutralize ela como fez o Propbank, com os papéis Arg0, Arg1, ….

- Uma boa meta: alcançar **90% de precisão na anotação**, pois os sistemas conseguirão sempre 10% menos.
  - Até não conseguir redefina as categorias exigindo menos detalhes

Agenda

- Projetos para o PB criados pelo NILC

- Questões em aberto desta área
  (1) Qual Corpus? Como conseguir um corpus balanceado para anotar? Quando o corpus é balanceado, representativo e ainda atual (não defasado)?
  (2) Como permanecer fiel à teoria? Como escrever um bom manual (não é trivial).
  (3) Que interfaces são melhores para cada tipo de problema e como garantir que elas não influenciam os resultados?
  (4) Quais as características importantes na seleção de anotadores? Como garantir que estão treinados de forma adequada?
  (5) Como criar um procedimento de anotação simples, rápido e confiável?
  (6) Como avaliar os resultados da anotação? Quais medidas de concordância são apropriadas?
  (7) Como armazenar os resultados?; Quando e para quem disponibilizar o corpus? Questões de licença, manutenção e distribuição.

Q3: Interface

- Como desenvolver uma boa interface?
  - Velocidade máxima!
    - Crie tarefas simples
    - Não use mouse, use ENTER
    - Customize a interface para vários projetos, mesmo que use projetos prontos
  - Evite construir uma interface tendenciosa (biased)
    - Cuidado com a ordem das escolhas das palavras
  - Evite mais do que 10 escolhas (regra 7 +-2)
  - Delimite uma região a ser anotada num contexto maior

Agenda

- Projetos para o PB criados pelo NILC

- Questões em aberto desta área
  (1) Qual Corpus? Como conseguir um corpus balanceado para anotar? Quando o corpus é balanceado, representativo e ainda atual (não defasado)?
  (2) Como permanecer fiel à teoria? Como escrever um bom manual (não é trivial).
  (3) Que interfaces são melhores para cada tipo de problema e como garantir que elas não influenciam os resultados?
  (4) Quais as características importantes na seleção de anotadores? Como garantir que estão treinados de forma adequada?
  (5) Como criar um procedimento de anotação simples, rápido e confiável?
  (6) Como avaliar os resultados da anotação? Quais medidas de concordância são apropriadas?
  (7) Como armazenar os resultados?; Quando e para quem disponibilizar o corpus? Questões de licença, manutenção e distribuição.

Q4: Anotadores

- Quanto treinar os anotadores? Nem muito nem pouco!

- **Treinar de menos:** Instruções vagas ou insuficientes. Resultado:
  - Anotadores criam um padrão próprio e divergem do gold standard.

- **Treinar de mais:** Se as instruções são longas, sem chances de interpretação, os anotadores acabam mecanizando a tarefa

Valorize seus anotadores

- Os anotadores são seu recurso mais valioso: eles (não você) conhecem os dados.

- Faça reuniões regulares:
  - Dê feedback regular; Diga que não há resposta corretas, mas que a sensibilidade deles ajuda a definir as respostas.
  - Incorpore seus comentários e sugestões no manual.

Agenda

- Projetos para o PB criados pelo NILC

- Questões em aberto desta área
  (1) Qual Corpus? Como conseguir um corpus balanceado para anotar? Quando o corpus é balanceado, representativo e ainda atual (não defasado)?
  (2) Como permanecer fiel à teoria? Como escrever um bom manual (não é trivial).
  (3) Que interfaces são melhores para cada tipo de problema e como garantir que elas não influenciam os resultados?
  (4) Quais as características importantes na seleção de anotadores? Como garantir que estão treinados de forma adequada?
  (5) Como criar um procedimento de anotação simples, rápido e confiável?
  (6) Como avaliar os resultados da anotação? Quais medidas de concordância são apropriadas?
  (7) Como armazenar os resultados?; Quando e para quem disponibilizar o corpus? Questões de licença, manutenção e distribuição.

Q5: Procedimento de Anotação

- Quando anotar várias variáveis, anote cada uma separadamente

- Permita anotadores discutir casos problemáticos

- Tenha um especialista para decidir casos difíceis
  - Super anotador: não vê as decisões dos anotadores
  - Juiz: vê as decisões dos anotadores

Heurísticas

- Faça as anotações simples primeiro.
- Peça que anotadores marquem o grau de certeza nas anotações,
    - pois para as que foram marcadas com grande certeza, deve haver alta taxa de concordância
- Avalie a estabilidade da anotação
- Crie um classificador com uma parte da anotação, anote o corpus e peça para anotadores revisarem

Agenda

- Projetos para o PB criados pelo NILC

- Questões em aberto desta área
  (1) Qual Corpus? Como conseguir um corpus balanceado para anotar? Quando o corpus é balanceado, representativo e ainda atual (não defasado)?
  (2) Como permanecer fiel à teoria? Como escrever um bom manual (não é trivial).
  (3) Que interfaces são melhores para cada tipo de problema e como garantir que elas não influenciam os resultados?
  (4) Quais as características importantes na seleção de anotadores? Como garantir que estão treinados de forma adequada?
  (5) Como criar um procedimento de anotação simples, rápido e confiável?
  (6) Como avaliar os resultados da anotação? Quais medidas de concordância são apropriadas?
  (7) Como armazenar os resultados?; Quando e para quem disponibilizar o corpus? Questões de licença, manutenção e distribuição.

Q6: O que medir?

- **O trabalho de anotação tem valor quando os anotadores concordam!**
- Mas o que medir?
  - Avalie **concordâncias individuais**, via kappa estatística (para vários anotadores: kappa estendido)
    - Quando o corpus não for balanceado, usar kappa não é uma solução boa. Neste caso use concordância simples
  - Avalie o **comportamento do grupo**
    - 10 anotadores, 20 categorias
    - Anotador 1 usa somente 3 categorias na metade dos exemplos, e ignora 30% das categorias: algo está errado!
  - Avalie **características do corpus**: balanceamento, partes mais difíceis

Kappa vs Concordância Simples

- Concordância Simples (precisão):

  A = número de escolhas que batem/número total de escolhas

- Como evitar concordância aleatória ?
  - Normalizar:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

- P(A) é a proporção de vezes que os anotadores concordam
- P(E) é a proporção de vezes que é esperado os juízes concordarem aleatoriamente

Kappa

Kappa index	Agreement
< 0.00	Less than chance
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

Landis, J.R.; & Koch, G.G. (1977). "The measurement of observer agreement for categorical data". *Biometrics* **33** (1): 159–174.

Agenda

- Projetos para o PB criados pelo NILC

- Questões em aberto desta área
  (1) Qual Corpus? Como conseguir um corpus balanceado para anotar? Quando o corpus é balanceado, representativo e ainda atual (não defasado)?
  (2) Como permanecer fiel à teoria? Como escrever um bom manual (não é trivial).
  (3) Que interfaces são melhores para cada tipo de problema e como garantir que elas não influenciam os resultados?
  (4) Quais as características importantes na seleção de anotadores? Como garantir que estão treinados de forma adequada?
  (5) Como criar um procedimento de anotação simples, rápido e confiável?
  (6) Como avaliar os resultados da anotação? Quais medidas de concordância são apropriadas?
  (7) Como armazenar os resultados?; Quando e para quem disponibilizar o corpus? Questões de licença, manutenção e distribuição.

Q7: Disponibilização

- Não basta anotar: questões técnicas devem ser tratadas:
  - Licença de Uso
  - Distribuição
  - Manutenção
  - Acrescentar novas anotações

Formatos de Intercâmbio de Dados

- Corpus ANC disponibiliza além de corpus, um padrão atual de intercâmbio: XCES, no formato GrAF (http://americannationalcorpus.org/)
- ANC2Go, saídas para Wordsmith, XML e outras
  - Metadados
  - Dados
  - Descreve todo o processo de anotação

Conclusão

- Anotação está se tornando uma ciência madura
  - Será necessário conhecer seus métodos
  - Há uma grande chance de unir dois tipos de pesquisadores em trabalhos conjuntos:
    - Linguistas de corpus
    - Linguistas computacionais

Vamos aproveitar a chance?

Aproveite a chance de se tornar este novo pesquisador que anota corpus.

Agradecimento

- Ed Hovy, pela sistematização do processo de anotação e pela palestra sobre Anotação de Corpus, no STIL 2011!

Obrigada pela atenção!