Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

ARS – Ferramenta de anotação de relações semânticas em textos escritos em português do Brasil

Helena de Medeiros Caseli
Leonardo Sameshima Taba

NILC-TR-13-03

Agosto, 2013

Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Este relatório apresenta uma descrição do uso da ferramenta Anotação de Relações Semânticas (ARS), desenvolvida com o propósito de auxiliar na tarefa de anotação manual de relações semânticas binárias em textos escritos em português do Brasil. A ferramenta enfoca sete relações, que são: is-a (hiponímia), property-of, part-of (meronímia), effect-of (causalidade), used-for, made-of e location-of. Essas relações são algumas das utilizadas no projeto de coleta de informações de senso comum *Open Mind Common Sense* do Brasil (OMCS-Br), e foram baseadas na teoria de representação de conhecimento de Minsky (1986). Além da descrição da ferramenta ARS e de seus processos de instalação e utilização, também são relatados experimentos e resultados do uso de técnicas de aprendizado de máquina (AM) para realizar a extração automática de relações semânticas em textos escritos na língua portuguesa do Brasil. Foram utilizados algoritmos de aprendizado de máquina supervisionados, de forma que textos com relações semânticas anotadas manualmente com a ferramenta foram utilizados como dados de treinamento. Os resultados mostram que a extração automática de relações semânticas usando métodos de AM tem bons resultados para relações com maior número de exemplos anotados, e que essa é uma direção promissora para pesquisas futuras sobre a extração de relações semânticas em textos escritos em português.

Índice

ARS – Ferramenta de anotação de relações semânticas em textos escritos em português do Brasil[1]

1 Introdução

Este relatório apresenta a ARS (Anotador de Relações Semânticas), uma ferramenta desenvolvida para auxiliar o processo de anotação de relações semânticas em textos escritos em português do Brasil. A ARS foi criada para dar suporte ao processo manual de anotação de relações semânticas derivadas da teoria de Minsky (1986) e está disponível na versão atual para anotação de 7 relações: is-a, part-of, made-of, property-of, effect-of, used-for e location-of.

Conforme apresentado em (TABA; CASELI, 2012), a ARS foi desenvolvida no contexto de um estudo sobre a extração automática de relações semânticas em textos escritos em português do Brasil (TABA, 2013), fornecendo suporte para a construção manual de um *corpus* anotado, recurso necessário para o treinamento de métodos de aprendizado de máquina (AM) supervisionados. Esses métodos são então utilizados para a extração automática das relações.

Atualmente, diversas aplicações do processamento de língua natural (PLN) como a tradução automática e sistemas de perguntas e respostas utilizam uma quantidade crescente de informações semânticas em seu processamento. No entanto, esse tipo de informação é custoso de se obter manualmente, já que demanda tempo e pessoal especializado. Algoritmos de AM supervisionados podem ser utilizados para extrair automaticamente esse tipo de informação, mas necessitam de dados de treinamento adequados; no caso da extração automática de relações semânticas, é necessário um conjunto de sentenças anotadas com as relações semânticas de interesse. A tarefa custosa de construção manual desse recurso motivou a criação da ARS, que surge como auxiliar no processo de construção manual de *corpora* anotados com relações semânticas.

O restante deste documento está organizado como segue. No capítulo 2 são apresentadas as regras gerais e específicas definidas para o processo de anotação de relações semânticas, com exemplos de cada uma. No capítulo 3 a ferramenta ARS é descrita em detalhes. No capítulo 4 a funcionalidade de extração automática de relações semânticas acoplada à ferramenta é explicada, e finalmente o capítulo 5 traz as conclusões deste relatório.

[1] Este trabalho foi desenvolvido com o apoio do CNPq e da FAPESP.

2 Anotação de relações semânticas

Este capítulo descreve as principais regras definidas para a marcação das 7 relações semânticas de interesse em textos escritos em português do Brasil. Essas relações semânticas são binárias e, portanto, dois termos[2] devem ser selecionados para marcá-las: t1 e t2.

As 7 relações de interesse são listadas a seguir:

1. is-a – Relação entre subclasse e superclasse ou definição de algo
   - t1: subclasse
   - t2: superclasse
2. part-of – Relação de parte e todo ou de membro de conjunto
   - t1: todo
   - t2: parte
3. made-of – Relação de produto e substância da qual é feito
   - t1: produto
   - t2: substância
4. property-of – Relação de característica de algo ou alguém
   - t1: algo/alguém
   - t2: característica
5. effect-of – Relação de causa e efeito
   - t1: algo
   - t2: consequência
6. used-for – Relação de ferramenta e utilidade
   - t1: ferramenta
   - t2: utilidade
7. location-of – Relação de localidade
   - t1: algo/alguém
   - t2: local

A ordem dos termos na relação faz parte de seu significado e deve seguir as convenções mostradas acima.

Assim, o processo de anotação das relações semânticas de interesse é dividido em duas etapas: delimitação dos termos envolvidos na relação (seção 2.1) e seleção da melhor relação (seção 2.2). Regras específicas para essas duas tarefas são apresentadas a seguir.

[2] Neste trabalho, um termo é definido como uma sequência de palavras com algum significado específico

2.1 Delimitação dos termos envolvidos na relação

A ARS pode identificar automaticamente substantivos, sequências de substantivos próprios e adjetivos como candidatos a termos, funcionalidade descrita em mais detalhes na Seção 4.3. No entanto, essa marcação automática prévia de candidatos é limitada e recomenda-se que seja revisada. Além disso, o anotador deve seguir algumas regras para a delimitação de termos, assegurando uma maior uniformidade na anotação dos dados e assim um melhor desempenho dos métodos computacionais que os utilizam. São elas:

1 – Priorizar termos menores

Sempre que possível, só deve ser inserida em um termo a quantidade de palavras necessária e suficiente para estabelecer a relação desejada. Se só o núcleo do SN for suficiente como termo da relação então devem ser removidos todos os artigos que acompanham o núcleo do SN; caso contrário, deve-se manter o SN completo.

Por exemplo:

- São Paulo enfrenta no campeonato dois adversários: Palmeiras e Vasco
  - is-a(Palmeiras, adversários)
  - is-a(Vasco, adversários)
- Primal=Scream fechou a noite em o palco principal
  - location-of(Primal=Scream, palco)
  - property-of(palco, principal)

2 – Não inserir números no termo

Por exemplo:

- dois adversários => adversários
- 70 igrejas => igrejas

3 – Não inserir aspas no termo

Se o termo estiver entre aspas, não se deve incluir as aspas no termo, mesmo que seja uma expressão, um ditado ou a fala de alguém.

4 – Não inserir o indicativo de "ordem" no termo

Por exemplo:

- última edição => última
  edição
- property-of(edição, última)

5 – Inserir o advérbio que modifica o adjetivo do termo

Por exemplo:

- o custo do dinheiro está muito alto
    - property-of(custo do dinheiro, muito alto)

6 – Inserir pronomes possessivos e demonstrativos no termo

Por exemplo:

- minha amiga
- esse espaço
- nossa meta
- seu filho
- seu bolso

7 – Não dividir termo específico (cargo, expressão fixa)

Se a aparente propriedade de um termo é, na verdade, uma especificação dele então não se deve dividir. Ex:

- taxa de juros
    - ~~property-of(taxa, de juros)~~
- gerente de comunicação empresarial de a Valmet
    - part-of(Valmet, gerente de comunicação empresarial)
- algo similar ocorre com: “deputado federal”, “previdência privada”, “ex-ministro da fazenda”, “região metropolitana”, “código de barras”, “área/região produtora de vinho”, “fatores meteorológicos”, “realidade virtual”, “meio interestelar”, etc.

Exceção

Essa regra se aplica a menos que seja necessário quebrar o termo para que ele se aplique a outras relações como em:

- os ministros de a Justiça, de a Defesa e de o Trabalho
    - property-of(ministros, de a Justiça)
    - property-of(ministros, de a Defesa)
    - property-of(ministros, de o Trabalho)

8 – Não dividir nome próprio (indicado pela presença de iniciais maiúsculas)

Se um termo genérico como "avenida", "colégio", "escola" aparecer com inicial maiúscula antes de um nome próprio ele é considerado como parte do nome próprio. Ex:

- Colégio Alberto=Levy
    - ~~is-a(Alberto=Levy, Colégio)~~

Exceção

Essa regra se aplica a menos que seja claramente impossível o termo genérico fazer parte do nome, como em:

- O Estado de Santa=Catarina
  - is-a(Santa=Catarina, Estado)

9 – Na dúvida, manter o maior termo

Se não for possível estabelecer a melhor delimitação de termos que mantenha o significado desejado, então deve-se manter o maior termo.

Por exemplo:

- A a tarde chegou a informação de que não seria possível antecipar o traslado , por=causa=de a autópsia .
  - effect-of(autópsia, não seria possível antecipar o traslado)
- O caminhão e algumas picapes 4 x4 levam gasolina , peças de reposição , comida , barracas etc .
  - location-of(barracas, caminhão e algumas picapes 4 x4)

Os termos envolvidos nas relações devem ser definidos seguindo-se essas regras. Assim, os candidatos a termos inicialmente marcados pela ferramenta ARS devem ser refinados seguindo as regras apresentadas nesta seção, sempre que necessário.

2.2 Seleção da relação mais adequada

Após a identificação dos termos presentes na sentença, o anotador deve selecionar a relação mais adequada para cada par de termos. Para tanto, algumas regras gerais também foram estabelecidas, são elas:

1 – Considerar a lei da transitividade

Principalmente nas relações is-a, location-of e part-of.

Por exemplo:

- location-of(Colégio Alberto=Levy, Indianópolis)
- is-a(Indianópolis, avenida)

ou

- is-a(Hamlet, peça)
- part-of(peça, personagem)

ou

- av. Indianópolis, Indianópolis, zona sul de São Paulo, SP
  - location-of(Indianópolis, Indianópolis)
  - location-of(Indianópolis, zona sul de São Paulo)

- location-of(zona sul de São Paulo, SP)
- part-of(Indianópolis, Indianópolis)
- part-of(zona sul de São Paulo, Indianópolis)
- part-of(SP, zona sul de São Paulo)

IMPORTANTE

- location-of prioriza o termo mais específico (Indianópolis) ao invés do mais geral (avenida)
- part-of prioriza o termo mais geral (peça) ao invés do mais específico (Hamlet)

  Exemplo que ilustra isso:
  - Linhares e Guriri , em o norte de o Espírito=Santo
    - location-of(Linhares, Espírito=Santo)
    - location-of(Guriri, Espírito=Santo)
    - part-of(Espírito=Santo, norte)

2 – Não considerar a polaridade da sentença para marcar as relações

Por exemplo:

- Mas este assunto não é pauta para as centrais sindicais
  - is-a(este assunto, pauta para as centrais sindicais)

Apesar da relação ideal ser a negação de is-a. considera-se, aqui, apenas o agrupamento, categorização, independente da polaridade.

3 – Não considerar a flexão verbal para marcar as relações

Para estabelecer uma relação com verbo deve-se sempre considerar a versão lematizada e na voz ativa do verbo presente no termo. Por exemplo:

- Em=parte , os exportadores têm sido ajudados por a desvalorização de o dólar , que torna seus produtos mais competitivos .
  - used-for(desvalorização de o dólar, ajudados)

  pois ao lematizar o verbo a relação faz sentido: used-for(desvalorização de o dólar, ajudar)
- A autópsia afirma que Jandira morreu devido=a a úlcera perfurada .
  - effect-of(úlcera perfurada, morreu)

  pois se lematizar o verbo a relação faz sentido: effect-of(úlcera perfurada, morrer)

A partir das regras gerais definidas nesta seção, a seleção da melhor relação semântica é realizada com base nas dicas apresentadas juntamente aos exemplos específicos de cada relação na próxima seção.

2.3 Exemplos

Para auxiliar na anotação da melhor relação para um determinado par de termos, a seguir são apresentados exemplos e dicas para cada uma das sete relações consideradas na versão atual da ARS: is-a, part-of, made-of, property-of, effect-of, used-for e location-of.

2.3.1 is-a

- t1: subclasse
- t2: superclasse

- <NOME> (<SIGLA>) ou SIGLA (<NOME>) ou TERMO (<DEFINIÇÃO>)

  is-a(<SIGLA>, <NOME>)

  - Bolsa de Mercadorias & Futuros (BM & F)
    - is-a(BM & F, Bolsa de Mercadorias & Futuros)
  - point average (a divisão de os pontos marcados por os pontos sofridos)
    - is-a(point average, divisão de os pontos marcados por os pontos sofridos)
- <DESIGNAÇÃO GENÉRICA> <NOME>

  is-a(<NOME>, <DESIGNAÇÃO GENÉRICA>)

  - o grupo Primal=Scream
    - is-a(Primal=Scream, grupo)
  - o ex-presidente FHC
    - is-a(FHC, ex-presidente)
  - minha=amiga Olga
    - is-a(Olga, minha amiga)
  - o seu filho Paulo
    - is-a(Paulo, seu filho)
  - o brasileiro Emerson=Fittipaldi
    - is-a(Emerson=Fittipaldi, brasileiro)

      Ver property-of para casos como "empresários brasileiros"
  - Reali que interpreta Ofélia
    - is-a(Reali, Ofélia)
- <DESIGNAÇÃO GENÉRICA> de <NOME>

  is-a(<NOME>, <DESIGNAÇÃO GENÉRICA>)

  - município de São=Paulo
    - is-a(São=Paulo, município)

2.3.2 part-of

- t1: todo
- t2: parte

DICA: usar part-of quando t2 for cargo de empresa/instituição/organização, procedimento ou local, ou seja, algo que possa ser dividido em partes.

- <PARTE> de <TODO>

  DICA: veja se é possível substituir "de <TODO>" por um adjetivo, se for, é property-of e não part-of

  - Hebron é uma cidade de a Cisjordânia
    - part-of(Cisjordânia, Hebron)
  - ao sul da Bahia
    - part-of(Bahia, sul)
  - A sala de o cinema
    - part-of(cinema, sala)
  - a capa de a revista
    - part-of(revista, capa)
  - a assinatura de o contrato
    - part-of(contrato, assinatura)
  - o cabelo de a peruca
    - part-of(peruca, cabelo)
  - o sangue de o corpo
    - part-of(corpo, sangue)
  - o sangue de o piloto (ou o sangue tirado de o piloto)
    - part-of(piloto, sangue)
  - o exército de radicais
    - part-of(exército, radicais)
  - O presidente de a empresa
    - part-of(empresa, presidente)
    - outro: presidente de a Fiesp => part-of(Fiesp, presidente)
  - apresentadora de TV
    - ~~part-of(apresentadora, de TV)~~

      Nesse caso part-of não se aplica pois TV é um termo muito genérico que não denota uma empresa específica, mas sim um coletivo de empresas e, portanto,

caracteriza um tipo específico de apresentadora . Nesse caso property-of é mais indicado.

- governo de Israel
  - ~~part-of(governo, de Israel)~~

    É possível substituir "de Israel" por um adjetivo (ex: israelense) e, por isso, é property-of e não part-of
- minha amiga de a Vila Madalena
  - ~~part-of(minha amiga, Vila Madalena)~~

    Veja que a amiga não é parte da Vila Madalena, ela pode ter apenas nascido ou morado um tempo por lá, mas nada indica que ainda esteja lá

- <PARTE> em <TODO> ou <PARTE> presente em <TODO>
  - DNA presente em o sangue
    - part-of(sangue, DNA)
- <PARTE> (<TODO>)
  - Indianápolis (zona sul de São Paulo)
    - part-of(zona sul de São Paulo, Indianápolis)
- Algo é parte de um procedimento
  - A concessionária não efetuou o conserto , porque a Fiat ainda não forneceu as peças de reposição .
    - part-of(conserto, peças de reposição)
- NÃO marcar part-of em casos como
  - o co-autor de a Teoria=da=Evolução
    - ~~part-of(Teoria=da=Evolução, co-autor)~~

      Algo semelhante ocorre com ~~parf-of(obra, autor)~~
  - regiões de clima subúmido
    - ~~part-of(regiões, clima)~~

2.3.3 made-of

- t1: produto
- t2: substância

  DICA: usar made-of quando t2 for a única ou a principal substância.

- Algo físico
  - o cabelo da peruca

- made-of(peruca, cabelo)

- o sangue de o corpo
  - ~~made-of(corpo, sangue)~~

    O corpo é formado por vários tecidos e o sangue é mais um deles, mas só com o sangue não se faz um corpo

- Aglomerado, conjunto
  - um casal de judeus
    - made-of(casal, judeus)
  - o exército de radicais
    - made-of(exército, radicais)
  - Os moradores de a região fizeram manifestações
    - made-of(manifestações, moradores)
  - amostra de palmeiras
    - made-of(amostra, palmeiras)
  - montanha de dados
    - made-of(montanha, dados)

2.3.4 property-of

- t1: algo/alguém
- t2: característica

DICA: usar property-of quando t2 for característica física (cor, peso, idade, tamanho, etc.), um adjetivo ou uma expressão preposicionada que indica característica (neste caso, geralmente, a expressão pode ser substituída por um adjetivo ou poderia se tal adjetivo existisse).

- <ALGO/ALGUÉM> de/a <CARACTERÍSTICA>
  - governo de Israel
    - property-of(governo, de Israel)

      Veja que aqui é possível substituir "de Israel" por um adjetivo: israelense
  - minha amiga de a Vila Madalena
    - property-of(minha amiga, de a Vila Madalena)

      Veja que aqui não existe um adjetivo que possa substituir "de a Vila Madalena", mas se ele existisse a substituição seria completamente aceitável como no caso anterior

- apresentadora de TV
  - property-of(apresentadora, de TV)

    TV é um termo muito genérico que não denota uma empresa específica, mas sim um coletivo de empresas e, portanto, caracteriza um tipo específico de apresentadora
- candidato a a presidência
  - property-of(candidato, a a presidência)
- o rei de a Espanha
  - property-of(rei, de a Espanha)
- a Copa do Mundo de os Estados Unidos
  - property-of(Copa do Mundo, de os Estados Unidos)
- O presidente de a República
  - ~~property-of(presidente, de a República)~~

    Ver part-of
- Hebron é uma cidade de a Cisjordânia
  - ~~property-of(Cisjordânia, Hebron)~~

    Ver part-of
- A sala de o cinema
  - ~~property-of(cinema, sala)~~

    Ver part-of
- a capa de a revista
  - ~~property-of(revista, capa)~~

    Ver part-of
- queda de circulação
  - ~~property-of(queda, de circulação)~~

    Não marcar property-of quando for o complemento e não uma característica

    Outros: oficinas para crianças, tipo de vegetação, ocupação humana (em "se esvai com a ocupação humana")
- FHC tem uma hérnia de disco
  - ~~property-of(FHC, hérnia)~~
  - property-of(hérnia, de disco)

    Algo semelhante ocorre em: property-of(artrite, reumatóide)
- a estante de Sônia
  - ~~property-of(estante, de Sônia)~~

Não marcar property-of quando for o agente ou o proprietário de algo

Outros: estada de Romário, declaração de Karadzic

- <ALGO/ALGUÉM> em <CARACTERÍSTICA>
  - banco oficial em o Rio=Grande=do=Sul
    - ~~property-of(banco, Rio=Grande=do=Sul)~~

      Não necessariamente é um banco gaúcho (de origem gaúcha)

      Neste caso location-of(banco, Rio=Grande=do=Sul) é mais adequado
- Característica física, nacionalidade
  - os empresários brasileiros
    - property-of(empresários, brasileiros)

      Ver is-a para casos como "o brasileiro Emerson Fittipaldi"
  - Ofélia faz anos amanhã
    - property-of(Ofélia, anos)

2.3.5 effect-of

- t1: algo
- t2: consequência

- Procedimento (autópsia, CPI) gera resultado, conclusão
  - resultado de a autópsia
    - effect-of(autópsia, resultado)

      Algo similar com "os resultados do trabalho": effect-of(trabalho, resultados)
  - resultado de a CPI
    - effect-of(CPI, resultado)
- Ação decorrente de algo

  DICA: tentar inserir apenas o verbo na ação e não toda a oração verbal, para que outras relações possam ser criadas entre termos que compõem a relação verbal.
  - o impacto humano tende a eliminar a vegetação natural
    - effect-of(impacto humano, eliminar)

2.3.6 used-for

- t1: ferramenta
- t2: utilidade

- Relações diretas
  - A concessionária não efetuou o conserto , porque a Fiat ainda não forneceu as peças de reposição .
    - used-for(peças de reposição, conserto)
- Liberdades poéticas
  - A o amanhecer de 25 de fevereiro , ele disse a a mulher que ia rezar em a Tumba de os Patriarcas , mas não levou seu manto de oração .
    - used-for(Tumba de os Patriarcas, rezar)

2.3.7 location-of

- t1: algo/alguém
- t2: local

DICA: usar location-of quando t2 for algo físico, palpável, ou quando for um local abstrato (p ex. "primeiro lugar", "vanguarda", etc.)

- <ALGO/ALGUÉM> em <LOCAL>
  - carreatas em a região metropolitana
    - location-of(carreatas, região metropolitana)
  - X está na vice-liderança de o campeonato
    - location-of(X, vice-liderança)
  - o brasileiro ficou em nono
    - location-of(brasileiro, nono)
  - Fernando=Collor assume o governo de Alagoas
    - location-of(Fernando=Collor, governo de Alagoas)
  - DNA presente em o sangue
    - location-of(DNA, sangue)
  - banco oficial em o Rio=Grande=do=Sul
    - location-of(banco, Rio=Grande=do=Sul)
- <ALGO> de <LOCAL>
  - Hebron é uma cidade de a Cisjordânia
    - location-of(Hebron, Cisjordânia)
  - A sala de o cinema
    - location-of(sala, cinema)
  - a capa de a revista

    - location-of(capa, revista)
  - o cabelo de a peruca
    - location-of(cabelo, peruca)
  - o sangue tirado de o piloto
    - ~~location-of(sangue, piloto)~~

      Não se aplica neste caso porque o sangue pode estar fora do piloto, no chão, por exemplo.
  - o exército de radicais
    - ~~location-of(radicais, exército)~~

      Não usa location-of neste caso porque o exército é um coletivo e não algo físico, palpável. Neste caso a relação made-of(exército, radicais) é indicada.
  - minha amiga de a Vila Madalena
    - ~~location-of(minha amiga, Vila Madalena)~~

      Veja que não se pode afirmar que a amiga está na Vila Madalena, ela pode ter apenas nascido ou morado um tempo por lá, mas nada indica que ainda esteja lá.
- <ALGO> (<LOCAL>)
  - Indianápolis (zona sul de São Paulo)
    - location-of(Indianápolis, zona sul de São Paulo)

3 Ferramenta de Anotação de Relações Semânticas (ARS)

A ferramenta de Anotação de Relações Semânticas (ARS) utilizada neste trabalho foi criada com o objetivo de auxiliar a anotação de textos em português com as relações semânticas descritas na seção anterior. Ela permite a marcação visual de termos e relações em sentenças, facilitando e acelerando o processo custoso de anotação manual. Além disso, a ARS também pode treinar e utilizar algoritmos de AM supervisionados que realizam a identificação automática de relações semânticas, funcionalidade descrita no próximo capítulo. Outras funcionalidades descritas nas próximas seções são a comparação de pacotes de sentenças anotados por dois anotadores distintos e a importação de textos puros e textos processados pelo *parser* PALAVRAS (BICK, 2000)

A ARS foi desenvolvida com a linguagem Java 1.6, usando a biblioteca de interface gráfica Swing e o ambiente de desenvolvimento integrado NetBeans 7.1. Também é um software *open-source* distribuído sob a GNU General Public License (GPL) versão 3. O motivo dessa escolha foi a inclusão da biblioteca do software WEKA 3.7.10 (HALL et al., 2009), também licenciado sob a GPLv3.

3.1 Requisitos e instalação

Os requisitos computacionais mínimos para executar a ferramenta são um computador com 256MB de RAM e processador de 300Mhz, além da máquina virtual Java versão 1.6. Pode ser utilizado qualquer sistema operacional onde a máquina virtual Java possa ser instalada (testes foram feitos sobre os sistemas Windows XP, Windows 7, Ubuntu 10.04 e Ubuntu 12.04).

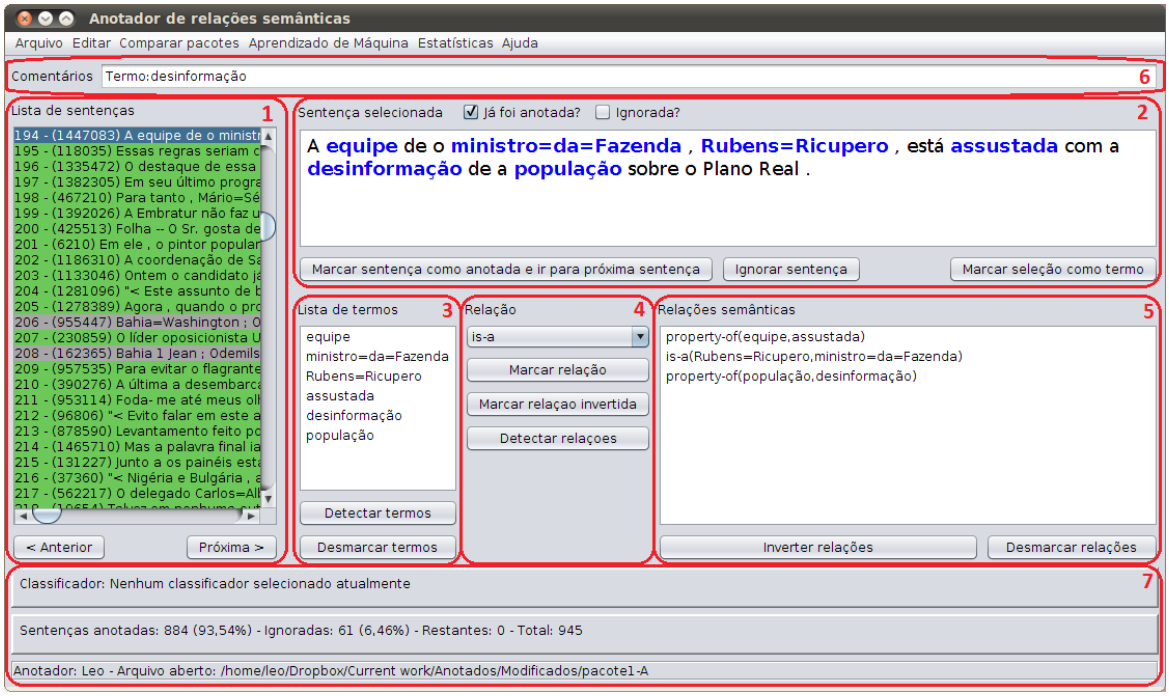
A ferramenta não necessita de instalação, basta fazer o *download* do pacote *jar* (ARS.jar) e executá-lo usando a máquina virtual Java. Isso pode ser feito através da linha de comando

```
java -jar ARS.jar
```

no diretório onde o pacote ARS.jar foi baixado, ou então clicando com o botão direito sobre o arquivo e selecionando a opção “Executar com Java” ou “Abrir com Java”, se seu sistema operacional associa os arquivos jar com o Java.

3.2 Tela principal

Figura 1 – Tela principal da ARS



Elementos da interface:

1. Lista de sentenças
2. Campo da sentença atualmente em edição
3. Lista de termos marcados
4. Relação a ser marcada
5. Lista de relações marcadas
6. Campo de comentários
7. Barras de notificações

3.3 Formato de entrada

A ferramenta trabalha com textos em um formato de entrada próprio baseado no padrão JSON[3] (*JavaScript Object Notation*). Nessa estrutura, *tokens*, termos e relações são tratados separadamente, cada qual como um objeto autocontido com seus próprios atributos. Esse formato foi escolhido sobre outros formatos tradicionais de codificação de *corpus* baseados

[3] http://www.json.org/

em XML como o Tiger XML (MENGEL; LEZIUS, 2000) e o XCES (IDE et al., 2000) por ser mais sucinto, mas igualmente portável e simples de se manipular computacionalmente.

O formato JSON se baseia em duas estruturas básicas, objetos e vetores. Objetos (definidos entre "{" e "}") são conjuntos de pares chave:valor, similares a vetores associativos, onde a chave é uma *string*; e vetores (definidos entre "[" e "]") são sequências ordenadas de valores. Valores podem ser *strings*, números, objetos, vetores, *true/false* (valores booleanos) ou *null* (valor inexistente). Baseada nessas estruturas, uma sentença é um objeto que contém outros objetos (*tokens*, termos, relações) e alguns campos (id, texto, entre outros). As estruturas definidas para a codificação de sentenças, *tokens*, termos e relações são mostradas nas Tabelas 1 a 4 a seguir.

Tabela 1 – Campos do objeto JSON que representam uma sentença

Campo	**Tipo**	**Descrição**
id	Numérico	Número que identifica a sentença
texto	String	Texto da sentença
comentários	String	Comentários adicionados pelos anotadores
anotada	Booleano	Se a sentença já foi anotada por algum anotador
ignorada	Booleano	Se a sentença foi marcada como ignorada (porque contém algum erro) pelos anotadores
tokens	Vetor de objetos	Vetor de objetos que representam cada *token* da sentença
termos	Vetor de objetos	Vetor de objetos que representam os termos marcados na sentença
relacoes	Vetor de objetos	Vetor de objetos que representam as relações semânticas marcadas na sentença
anotadores	Vetor de strings	Vetor de strings que armazenam o nome dos anotadores que trabalharam na sentença

Tabela 2 – Campos do objeto JSON que representam um *token*

Campo	**Tipo**	**Descrição**
t	String	Forma superficial do *token*
l	String	Lema (forma base) do *token*
pos	String	Etiquetas *part-of-speech* do *token* (cf. anotado pelo PALAVRAS (BICK, 2000))
sin	String	Papel sintático do *token* na árvore de dependências da sentença (cf. anotado pelo PALAVRAS (BICK, 2000))

Tabela 3 – Campos do objeto JSON que representam um termo

Campo	**Tipo**	**Descrição**
de	Numérico	Índice do *token* (no vetor de *tokens* da sentença) onde o termo se inicia)
para	Numérico	Índice do *token* (no vetor de *tokens* da sentença) onde o termo termina)

Tabela 4 – Campos do objeto JSON que representam uma relação

Campo	**Tipo**	**Descrição**
r	String	Qual relação semântica está marcada (is-a, property-of, etc.)
t1	Numérico	Índice do termo que é o primeiro participante da relação
t2	Numérico	Índice do termo que é o segundo participante da relação

A seguir é mostrado um exemplo completo de uma sentença codificada nesse formato, retirada dos pacotes anotados do *corpus* CETENFolha.


```
{
    "id": 271463,
    "texto": "Em a virada para o real , os valores foram convertidos para a
nova moeda , mas sem considerar a desvalorização de o dólar .",
    "comentarios": "Termo:desvalorização=de=o=dólar",
    "anotada": true,
    "ignorada": false,
    "tokens": [
        { "t": "Em", "l": "em", "pos": "<sam-> PRP", "sin": "@ADVL>" },
        { "t": "a", "l": "o", "pos": "<-sam> <artd> DET F S", "sin":
"@>N" },
        { "t": "virada", "l": "virada", "pos": "N F S", "sin": "@P<" },
        { "t": "para", "l": "para", "pos": "PRP", "sin": "@N<" },
        { "t": "o", "l": "o", "pos": "<artd> DET M S", "sin": "@>N" },
        { "t": "real", "l": "real", "pos": "N M S", "sin": "@P<" },
        { "t": ",", "l": ",", "pos": null, "sin": null },
        { "t": "os", "l": "o", "pos": "<artd> DET M P", "sin": "@>N" },
        { "t": "valores", "l": "valor", "pos": "N M P", "sin": "@SUBJ>" },
        { "t": "foram", "l": "ser", "pos": "<fmc> V PS/MQP 3P IND VFIN",
"sin": "@FAUX" },
        { "t": "convertidos", "l": "converter", "pos": "V PCP M P", "sin":
"@IMV @#ICL-AUX<" },
        { "t": "para", "l": "para", "pos": "PRP", "sin": "@<ADVL" },
        { "t": "a", "l": "o", "pos": "<artd> DET F S", "sin": "@>N" },
```

```
        { "t": "nova", "l": "novo", "pos": "ADJ F S", "sin": "@>N" },
        { "t": "moeda", "l": "moeda", "pos": "N F S", "sin": "@P<" },
        { "t": ",", "l": ",", "pos": null, "sin": null },
        { "t": "mas", "l": "mas", "pos": "<co-advl> KC", "sin": "@CO" },
        { "t": "sem", "l": "sem", "pos": "PRP", "sin": "@<ADVL" },
        { "t": "considerar", "l": "considerar", "pos": "V INF", "sin":
"@IMV @#ICL-P<" },
        { "t": "a", "l": "o", "pos": "<artd> DET F S", "sin": "@>N" },
        { "t": "desvalorização", "l": "desvalorização", "pos": "N F S",
"sin": "@<ACC" },
        { "t": "de", "l": "de", "pos": "<sam-> PRP", "sin": "@N<" },
        { "t": "o", "l": "o", "pos": "<-sam> <artd> DET M S", "sin":
"@>N" },
        { "t": "dólar", "l": "dólar", "pos": "N M S", "sin": "@P<" },
        { "t": ".", "l": ".", "pos": null, "sin": null }
    ],
    "termos": [
        { "de": 13, "ate": 13 },
        { "de": 14, "ate": 14 }
    ],
    "relacoes": [
        { "r": "property-of", "t1": 1, "t2": 0 }
    ],
    "anotadores": [
        "Leo"
    ]
}
```


Essa sentença tem dois termos marcados (“nova” e “moeda”) e a relação property-of entre eles.

Um arquivo de entrada da ferramenta é composto por duas partes: um cabeçalho que pode conter quaisquer informações, já que é ignorado pela ferramenta; e uma sequência de sentenças codificadas no formato mostrado, com uma sentença por linha. O cabeçalho é terminado por uma linha contendo apenas “===” (3 caracteres “=”).

3.4 Conversão de textos para o formato JSON

Como citado anteriormente, a ferramenta aceita como entrada conjuntos de sentenças codificadas no formato JSON apresentado. Textos processados pelo *parser* PALAVRAS (BICK, 2000) podem ser transformados pela ferramenta para o formato apresentado, bastando selecionar a opção “Converter arquivo processado pelo PALAVRAS” dentro do submenu

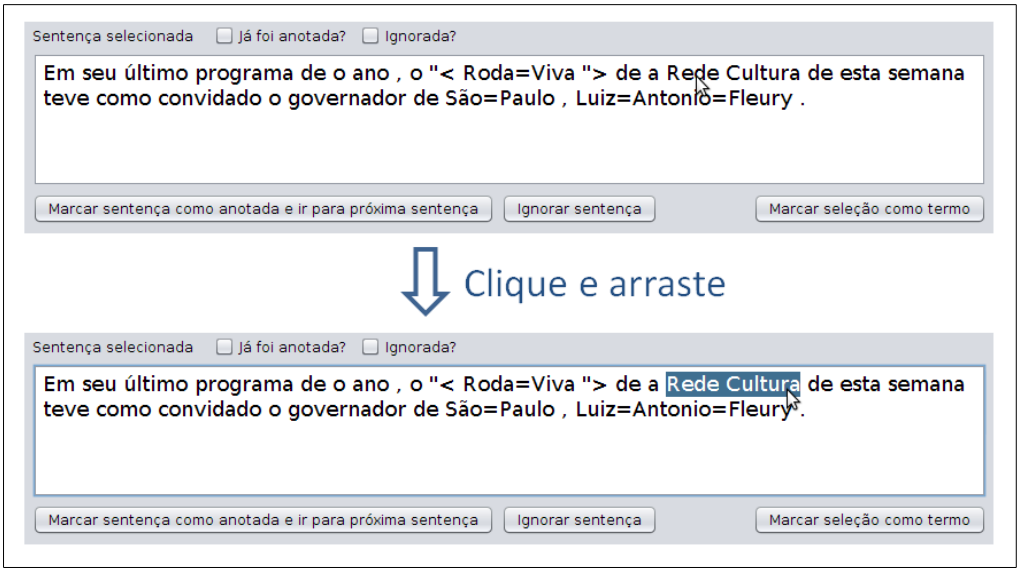
"Importar" no menu "Arquivo" e selecionar o arquivo apropriado. Após selecionar o arquivo processado pelo PALAVRAS, é necessário escolher o nome e o local do arquivo JSON que será gerado. Após a transformação o arquivo pode ser aberto selecionando a opção "Abrir" no menu "Arquivo". O arquivo processado pelo PALAVRAS deve ter a mesma formatação do arquivo texto_exemplo.palavras fornecido junto com o pacote da ferramenta.

Podem ser utilizados também textos puros como entrada, ou seja, textos sem anotações morfossintáticas. Esses arquivos podem ser transformados para o formato JSON selecionando a opção "Converter arquivo de texto puro" dentro do submenu "Importar" no menu "Arquivo". No entanto, é importante notar que nesse caso a anotação automática de relações, descrita no capítulo 4, não poderá ser aplicada, já que as principais informações utilizadas pelos métodos automáticos de anotação vêm das informações morfossintáticas adicionais. Os arquivos de texto devem estar na codificação UTF-8 e com uma sentença por linha.

3.5 Anotação de termos

Inicialmente uma sentença deve ser selecionada na lista de sentenças à esquerda. Feito isso, a sentença será mostrada no campo de edição principal. Para realizar a marcação de termos, basta selecionar um ou mais *tokens* na sentença com o mouse, clicando no termo inicial e arrastando o mouse até o *token* final do termo (Figura 2). Com os *tokens* selecionados, pode-se clicar com o botão direito e selecionar a opção "Marcar seleção como termo" do menu de contexto ou então clicar no botão "Marcar seleção como termo". Dessa forma o conjunto de *tokens* será marcado como termo e aparecerá em negrito e azul, além de aparecer na lista de termos.

Figura 2 – Seleção de *tokens* para marcação de termos



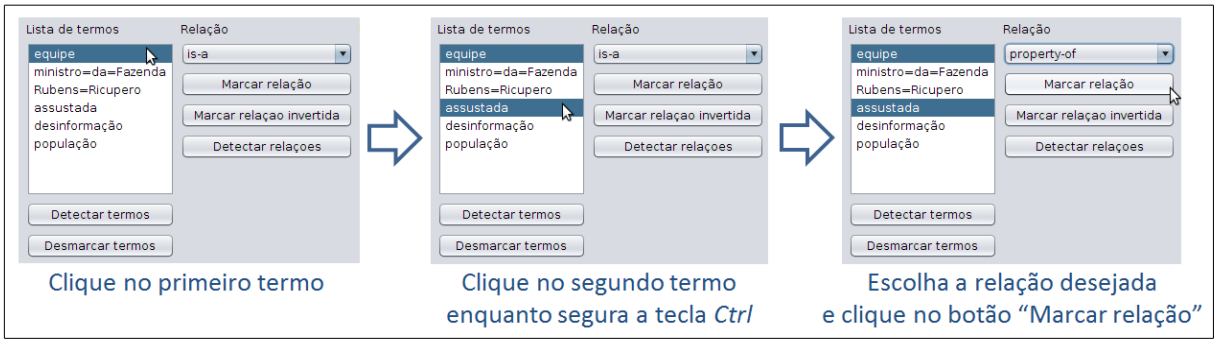
A desmarcação de termos pode ser feita de forma semelhante: pode-se clicar com o botão direito sobre um termo no campo de edição da sentença e escolher a opção "Desmarcar termo" do menu de contexto, ou então escolher o termo desejado na lista de termos e clicar no botão "Desmarcar termo".

É importante notar que um termo não pode conter outro termo; a ferramenta impede a marcação de um termo que englobe parte de ou todo outro termo.

3.6 Anotação de relações

Para marcar uma relação semântica entre um par de termos, basta selecionar os dois termos desejados na lista de termos anotados (clique no primeiro termo e, enquanto mantém a tecla *Ctrl* do teclado pressionada, clique no segundo termo), escolher a relação desejada no *Combobox* de relações possíveis e pressionar o botão "Marcar relação". Feito isso, a relação será marcada e aparecerá na lista de relações (Figura 3).

Figura 3 – Marcação de relação semântica entre termos



A ordem dos termos na relação marcada dessa forma é a mesma ordem dos termos na sentença, ou seja, o termo que aparece antes estará na posição de primeiro argumento da relação e o termo que aparece depois estará na posição de segundo argumento. Se desejar marcar a relação na ordem inversa, basta selecionar os termos desejados e clicar no botão "Marcar relação invertida". Se quiser inverter a ordem de uma relação já marcada, basta dar um clique duplo sobre a relação desejada na lista de relações, ou então selecionar a relação e clicar no botão "Inverter relações" ou clicar com o botão direito e selecionar a opção "Inverter relação" do menu de contexto.

A desmarcação de relações é feita de forma semelhante: basta selecionar a relação na lista de relações e clicar no botão "Desmarcar relações", ou então selecionar a relação, clicar com o botão direito e selecionar a opção "Desmarcar relação" do menu de contexto. Pode-se

selecionar mais de uma relação (mantendo a tecla *Ctrl* pressionada) e desmarcar ou inverter todas de uma vez, clicando nos botões "Desmarcar relações" ou "Inverter relações".

Podem ser marcadas diferentes relações com os mesmos argumentos, ou seja, dados dois termos A e B as relações is-a(A, B), property-of(A, B), etc. podem ser marcadas na mesma sentença. No entanto, uma mesma relação com argumentos invertidos, por exemplo uma instância is-a(A, B) e uma is-a(B, A), não pode ser marcada na mesma sentença.

Note também que quando um termo que faz parte de alguma relação é desmarcado, todas as relações em que esse termo participa também são desmarcadas automaticamente.

3.7 Verificação de concordância entre anotadores

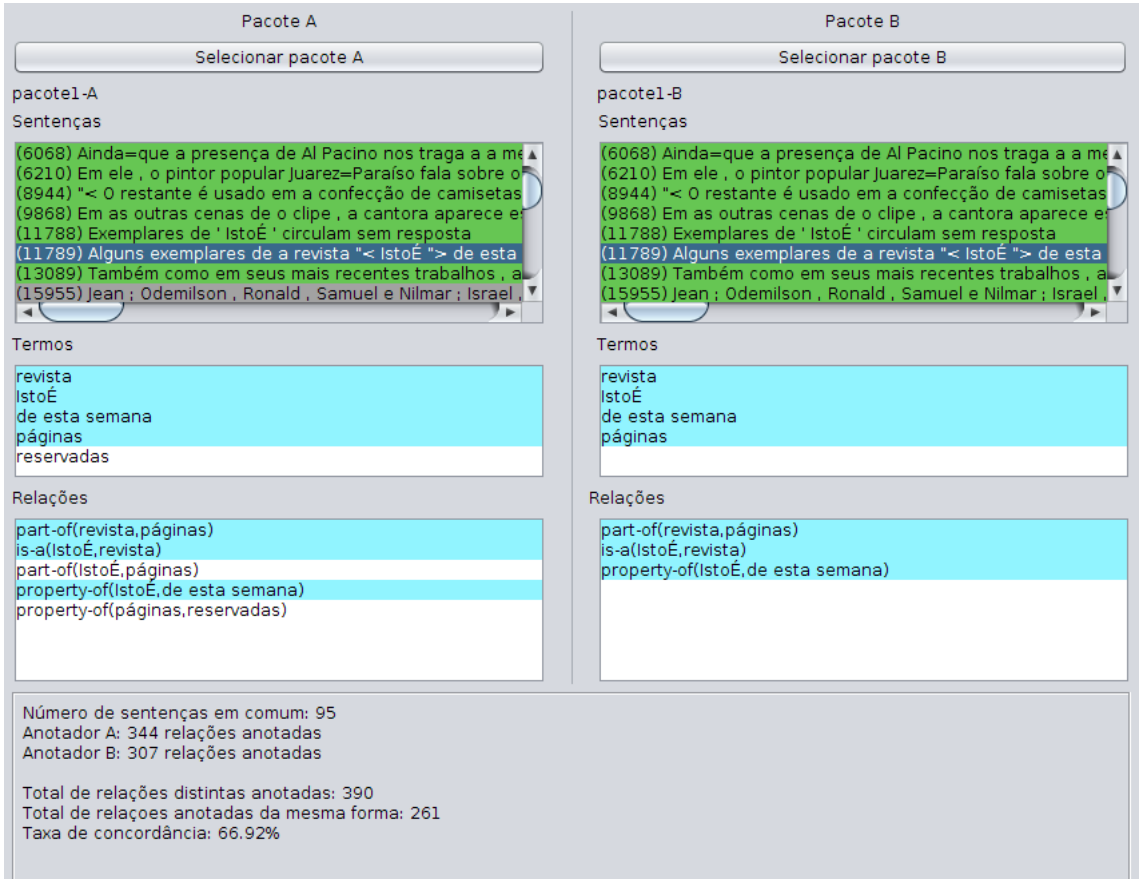
Uma das características que diferenciam esta ferramenta das demais é a possibilidade de comparar a anotação feita por dois anotadores distintos, permitindo assim o cálculo da concordância entre suas anotações. A comparação pode ser feita entre dois pacotes de sentenças que tenham alguma intersecção entre si. Por exemplo, no trabalho de mestrado em que esta ferramenta foi desenvolvida, cada anotador foi responsável por anotar dois pacotes com cerca de 1000 sentenças cada um, sendo que haviam 100 sentenças em comum entre esses pacotes. Portanto, foram anotadas cerca de 3800 sentenças, com 200 sentenças tendo sido anotadas por ambos os anotadores. A Figura 4 ilustra a interface de comparação entre pacotes.

O cálculo da concordância é feito sobre a intersecção dos pacotes, ou seja, sobre as sentenças em comum. A concordância é definida conforme a razão:

$$\frac{\textit{Número de relações anotadas de forma idêntica}}{\textit{Número de relações distintas anotadas}}$$

Por exemplo, se considerarmos apenas uma sentença em comum entre dois pacotes, e o anotador A marcou nessa sentença as relações "is-a(A, B)" e "is-a(B, C)", e o anotador B as relações "is-a(A, B)" e "part-of(C, B)", temos 1 relação marcada de forma idêntica (is-a(A, B)) e 3 relações distintas no total. Portanto, a taxa de concordância é de 33,3%.

Figura 4 – Interface de comparação de pacotes[4]



[4] Termos e relações destacados em azul foram marcados da mesma forma por ambos os anotadores

4 Anotação automática de relações semânticas

O capítulo anterior descreveu o funcionamento geral da ARS como ferramenta de auxílio ao processo manual de anotação de relações semânticas, tornando esse processo mais automatizado e rápido. Este capítulo apresenta a funcionalidade de anotação automática de relações semânticas acoplada à versão mais recente da ARS.

Para tanto, algoritmos de AM foram treinados a partir de *corpora* anotados usando a ARS como descrito no capítulo 3. Nesse processo de anotação semiautomática a ARS foi utilizada para anotar um subconjunto de dois *corpora* distintos: CETENFolha[5] e FAPESP (AZIZ; SPECIA, 2011). O processo de preparação dos *corpora* usados no treinamento dos algoritmos de AM é descrito na próxima seção.

4.1 *Corpora* de Treinamento e Teste

Os algoritmos de AM utilizados na ferramenta são do tipo supervisionado, ou seja, são treinados sobre um conjunto de dados anotados com informações de interesse – neste caso, sentenças anotadas com seus termos e relações semânticas entre eles. Após o treinamento os classificadores podem ser utilizados para encontrar automaticamente novas relações semânticas em textos.

Inicialmente o *corpus* CETENFolha foi escolhido como *corpus* de treinamento devido a sua abrangência de temas e tamanho. O CETENFolha é composto por textos do jornal Folha de São Paulo do ano de 1994, processados morfossintaticamente pelo *parser* PALAVRAS (BICK, 2000). Um subconjunto de cerca de 3700 sentenças foi selecionado e anotado com seus termos e relações semânticas de interesse por 2 anotadores. A taxa de concordância entre eles, calculada conforme descrito na Seção 3.7, foi de 69,15%.

Posteriormente foi realizada a anotação de cerca de 500 sentenças do *corpus* FAPESP (AZIZ; SPECIA, 2011), composto por textos de divulgação científica da revista Pesquisa FAPESP também processados morfossintaticamente pelo *parser* PALAVRAS. Essa anotação foi feita por apenas um dos anotadores já que, após a marcação das sentenças do CETENFolha, ambos os anotadores obtiveram uma taxa de concordância aceitável, não sendo mais necessária a anotação em paralelo. Esta porção anotada do *corpus* FAPESP foi utilizada como conjunto de teste em uma das avaliações dos algoritmos de AM, conforme descrito na próxima seção.

A Tabela 5 mostra a quantidade de sentenças que foram anotadas em cada *corpus* e o número de *tokens* contidos em cada um e a Tabela 6 traz o número de instâncias de cada

[5] http://www.linguateca.pt/cetenfolha/

relação marcadas manualmente pelos anotadores em cada *corpus*. A tabela 6 contém também o número de instâncias negativas (pares de termos entre os quais não existe relação) – elas são importantes para que os métodos de AM saibam reconhecer os pares de termos que não são relacionados.

As instâncias negativas são geradas automaticamente pela ARS a partir de todos os pares de termos marcados em sentenças mas entre os quais não existe nenhuma relação. Uma observação importante é que são considerados apenas os termos que fazem parte de alguma relação, sendo que termos marcados e não participantes em nenhuma relação não são considerados. Por exemplo, se em uma sentença foram marcados os termos A, B, C e D e as relações is-a(A, B) e property-of(B, C), serão geradas as instâncias negativas none(A, C) e none(C, A), indicando que não existe relação entre os termos A e C. O termo D é ignorado pois não faz parte de nenhuma relação positiva. Isso é feito para diminuir o problema de desbalanceamento da classe negativa, que é muito maior que as demais classes positivas.

Finalmente, a Tabela 7 traz alguns exemplos das anotações realizadas.

Tabela 5 – Descrição dos subconjuntos dos *corpora* anotados com o auxílio da ARS.

Corpus	***Tokens***	**Sentenças**
CETENFolha	96803	3679
FAPESP	13662	500
Total	**110465**	**4179**

Tabela 6 – Número de instâncias de cada relação anotadas manualmente nos *corpora* mais as instâncias negativas geradas automaticamente

Relação	**CETENFolha**	**FAPESP**
property-of	4909	853
is-a	2816	367
part-of	1976	306
location-of	1683	248
effect-of	163	84
used-for	136	68
made-of	74	60
Subtotal	**11757**	**1986**
Classe negativa	**104135**	**23985**
Total	**115892**	**25971**

Tabela 7 – Exemplos de sentenças do *corpus* de treinamento anotadas manualmente com relações semânticas de interesse usando a ARS.

Sentença	Relações
Os dentes do mais antigo orangotango	parf-of(orangotango, dentes) property-of(orangotango, mais antigo)
Outras empresas, como a Shorts, produtora de motores aeronáuticos em Belfast (Irlanda) e a Perkins deram seus primeiros passos em a mesma direção.	is-a(Shorts, empresas) location-of(Shorts, Belfast) location-of(Belfast, Irlanda) property-of(motores, aeronáuticos) is-a(Perkins, empresas)
(…) eventuais perdas operacionais com a desvalorização de o dólar em relação a o real estão (...)	property-of(perdas, eventuais) property-of(perdas, operacionais) effect-of(desvalorização de o dólar, perdas)
Os fãs de " Carruagens de Fogo " vão reconhecer os painéis de madeira escura (...)	made-of(painéis, madeira) property-of(madeira, escura)
Os aminoácidos são usados em a " construção " de os músculos	used-for(aminoácidos, construção)

4.2 Treinamento e Teste dos algoritmos de AM

A partir dos *corpora* anotados descritos na seção 4.1, dois algoritmos de AM foram treinados: árvores de decisão C4.5 (QUINLAN, 1993) e S*upport Vector Machines* (SVMs) (VAPNIK, 1995).

As árvores de decisão são um dos algoritmos mais utilizados em estudos com AM devido à sua simplicidade e bons resultados. Seu funcionamento se baseia na criação de árvores onde cada nodo representa uma tomada de decisão sobre uma variável da representação do problema, com as folhas da árvore indicando a classificação de uma dada instância. A implementação de árvore de decisão C4.5 escolhida para realizar os experimentos foi a J48 do pacote WEKA versão 3.7.10 (HALL et al., 2009) com o parâmetro $C^6 = 0{,}25$.

O segundo algoritmo de AM escolhido foram as *Support Vector Machines* (SVMs) (VAPNIK, 1995), classificadores binários que figuram no estado da arte entre os algoritmos de AM devido ao seu bom desempenho. Basicamente, as SVMs utilizam propriedades geométricas para encontrar um hiperplano que melhor separe dois conjuntos de dados; esse hiperplano é encontrado através da resolução de um problema de otimização quadrático.

[6] Fator de confiança na poda da árvore. Quanto menor, mais podas serão feitas.

A implementação de SVM escolhida foi a do software SVM Light (JOACHIMS, 1998), com *kernel* polinomial de grau 3 e parâmetro C[7] = 0,01. Preferiu-se o uso do SVM Light ao invés da implementação do Weka pois os testes com o último tomavam cerca de 5 vezes mais tempo que com o primeiro.

Esses dois algoritmos foram treinados e testados com os dados anotados do *corpus* CETENFolha, Os resultados da realização de *10-fold cross-validation* para cada algoritmo estão resumidos na Tabela 8. Para realizar a *featurização* dos dados foram definidas 288 *features* de diferentes níveis linguísticos, descritas em (TABA, 2013).

Tabela 8 – Resultados do *10-fold cross-validation* para ambos os classificadores treinados sobre o *corpus* CETENFolha

	Árvore de decisão			**SVM**		
Relação	**Precisão**	**Cobertura**	**Medida-F**	**Precisão**	**Cobertura**	**Medida-F**
property-of	90,5%	80,0%	84,9%	89,6%	81,6%	85,4%
is-a	76,9%	56,3%	65,0%	78,2%	65,1%	71,0%
part-of	66,4%	37,2%	47,7%	56,9%	41,4%	47,9%
location-of	62,2%	21,4%	31,8%	51,8%	27,8%	36,2%
effect-of	34,0%	10,4%	16,0%	45,5%	16,9%	24,6%
made-of	33,3%	2,8%	5,2%	58,2%	24,3%	34,3%
used-for	17,5%	5,1%	8,0%	50,8%	17,7%	26,2%
Média	**54,4%**	**30,4%**	**39,0%**	**61,6%**	**39,2%**	**47,9%**

Com o objetivo de avaliar a capacidade de generalização desses classificadores quando apresentados a dados não vistos no treinamento, foi realizado um teste sobre a porção anotada do *corpus* FAPESP. Assim, ambos os classificadores foram treinados sobre o *corpus* CETENFolha e avaliados sobre o *corpus* FAPESP. Os resultados desse experimento, descritos na Tabela 9, mostram que os classificadores tem boa capacidade de generalização e que, portanto, podem ser utilizados para a descoberta de novas relações semânticas em textos inéditos. Mais informações sobre os experimentos realizados e resultados estão descritos em (TABA, 2013).

[7] Fator de compensação entre os erros no treinamento e a margem dos vetores de suporte.

Tabela 9 – Resultados dos experimentos dos classificadores treinados com o *corpus* CETENFolha e testados sobre o *corpus* FAPESP

	Árvore de decisão			**SVM**		
Relação	**Precisão**	**Cobertura**	**Medida-F**	**Precisão**	**Cobertura**	**Medida-F**
property-of	89,1%	83,3%	86,1%	91,0%	84,3%	87,5%
is-a	60,0%	26,2%	36,4%	57,6%	32,1%	41,2%
part-of	59,5%	35,0%	44,1%	52,1%	37,1%	43,3%
location-of	39,1%	10,9%	17,0%	39,6%	17,7%	24,5%
effect-of	40,0%	7,1%	12,1%	50,0%	8,3%	14,3%
made-of	0,0%	0,0%	0,0%	36,4%	7,3%	12,1%
used-for	38,5%	7,4%	12,3%	40,9%	13,2%	20,0%
Média	**23,3%**	**22,8%**	**23,0%**	**52,5%**	**28,6%**	**37,0%**

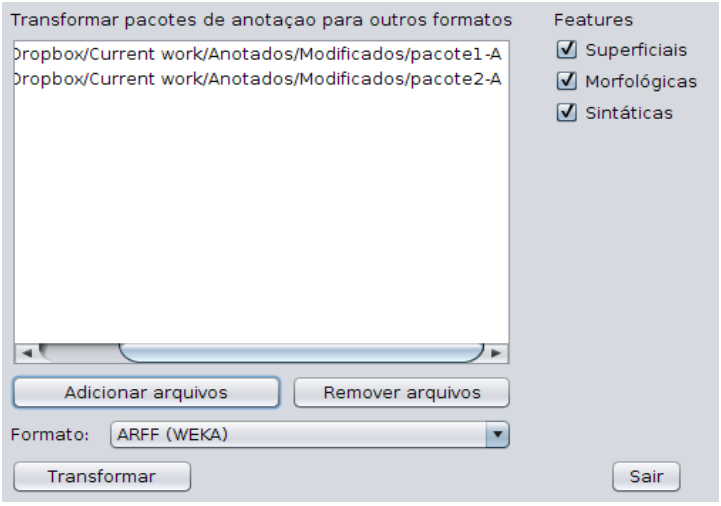
4.3 Anotação automática de relações semânticas na ARS

Ambos os algoritmos de AM apresentados na seção anterior podem ser acoplados à ARS, permitindo assim o seu treinamento e posterior utilização na marcação automática de relações semânticas.

O algoritmo de árvore de decisão J48 já vem integrado à ferramenta. O classificador SVM não vem integrado mas pode ser baixado do site http://svmlight.joachims.org/. Após baixado, ele pode ser acoplado à ARS no menu "Aprendizado de Máquina", item "Configurações", e selecionando o caminho dos programas "svm_learn" e "svm_classify". A versão do SVM Light utilizada nos testes descritos neste relatório foi a 6.02.

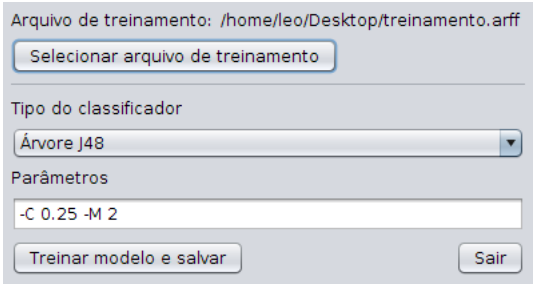
Para utilizar os classificadores em conjunto com a ferramenta, primeiramente eles precisam ser treinados com sentenças etiquetadas corretamente. Para fazer isso, primeiramente os pacotes contendo as sentenças com termos e relações manualmente anotadas devem ser transformados para o formato adequado a cada classificador (ARFF, no caso das árvores de decisão, ou SVM Light), funcionalidade encontrada no menu "Aprendizado de Máquina", opção "Transformar pacotes para outros formatos". Basta selecionar os arquivos que contêm os conjuntos de sentenças anotadas e clicar no botão "Transformar" (Figura 5). As *features* que serão utilizadas na transformação dos dados podem ser modificadas caso necessário, mas, conforme os resultados mostrados em (TABA, 2013), o uso dos três conjuntos de *features* traz melhores resultados para os classificadores.

Figura 5 – Tela de transformação de pacotes de sentenças anotadas para conjuntos de dados de treinamento



Após a transformação dos pacotes de sentenças, o treinamento dos classificadores pode ser feito através do menu "Aprendizado de Máquina", opção "Treinar classificador". Nessa tela (Figura 6) o arquivo de treinamento criado anteriormente deve ser selecionado. Em seguida, escolhe-se o classificador desejado e clica-se no botão "Treinar". A última etapa é escolher o local onde se deseja salvar o modelo do classificador treinado. Feito isso, a ferramenta iniciará o treinamento do classificador escolhido e avisará quando o processo terminar. É importante frisar que o tamanho do arquivo de treinamento influencia diretamente o tempo de treinamento dos classificadores, que pode variar de alguns minutos até algumas horas, e enquanto o treinamento estiver em execução a ferramenta não realizará outras tarefas.

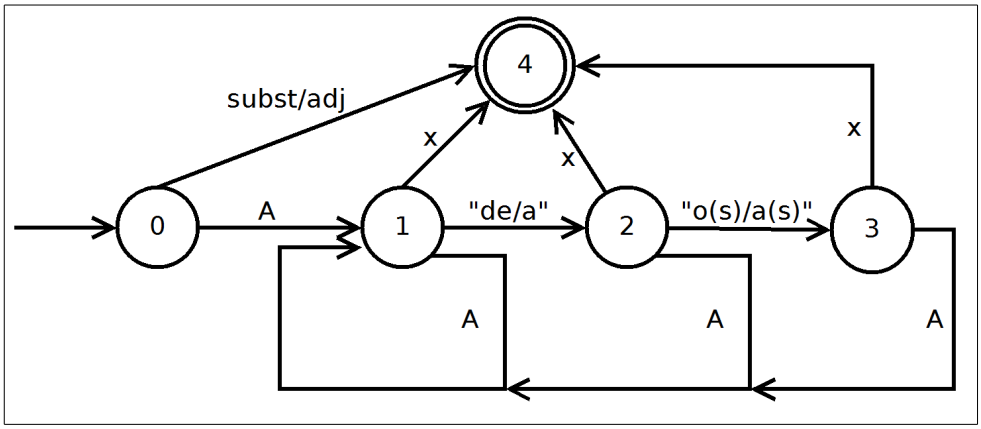
Figura 6 – Tela de treinamento de um classificador



Antes de aplicar o classificador treinado é necessário realizar a anotação dos termos nas sentenças, já que a tarefa de extração de relações semânticas consiste em duas etapas, a identificação de termos e a identificação de relações entre eles. Visando auxiliar a primeira etapa, a ARS tem uma funcionalidade limitada de detecção automática de termos. Para utilizá-

la basta clicar no botão “Detectar termos”, localizado abaixo da lista de termos marcados, ou ir no menu “Aprendizado de Máquina” e clicar na opção “Detectar termos em todas as sentenças deste pacote”. Essa detecção é limitada e deve ser revisada pelo anotador. Basicamente, ela é feita utilizando um autômato finito (Figura 7) que reconhece substantivos, adjetivos e sequências de substantivos próprios como termos.

Figura 7 – Autômato finito que reconhece termos[8]



Depois que os termos em uma sentença estejam anotados o classificador treinado pode ser aplicado. Primeiro, ele precisa ser selecionado (menu “Aprendizado de Máquina”, opção “Selecionar classificador”), e então basta clicar no botão “Detectar relações” na interface principal, ou então na opção “Detectar relações em todas as sentenças deste pacote” no menu “Aprendizado de Máquina”. Assim como no treinamento a detecção automática de relações pode levar vários minutos em cada sentença, especialmente utilizando classificadores SVM.

De forma a simplificar o trabalho de futuros anotadores, um classificador treinado de árvore de decisão e um SVM são fornecidos junto com o pacote da ferramenta ARS (arquivos classificador_cetenfolha.j48 e classificador_cetenfolha.svm), permitindo assim o uso imediato da funcionalidade de detecção automática de relações semânticas. É importante notar que os classificadores fornecidos foram treinados sobre o *corpus* CETENFolha, anotado conforme descrito na Seção 4.1.

[8] No autômato o símbolo “A” indica um *token* iniciado com letra maiúscula, “subst/adj” indica um *token* que é substantivo ou adjetivo e o símbolo “x” indica qualquer *token* que não participe de alguma transição de estado. Quando o estado 4 é alcançado o termo identificado é marcado. No estado 0 se um *token* é iniciado com letra maiúscula e é um subtantivo ou adjetivo, ele segue para o estado 1

5 Conclusão

Esse relatório apresentou a ferramenta de anotação de relações semânticas ARS, desenvolvida durante o trabalho de mestrado descrito em (TABA, 2013), bem como as regras definidas para guiar este processo. A ARS pode ser usada tanto para anotar textos sem o auxílio de um modelo de AM previamente treinado (anotação manual descrita no capítulo 3) quanto com o auxílio de tal modelo (anotação automática conforme descrito no capítulo 4).

A língua portuguesa ainda carece de recursos e ferramentas de qualidade para a realização de estudos em PLN, particularmente em nível semântico. Assim, a ARS busca facilitar a criação de *corpora* anotados com informações semânticas através da possibilidade de identificação automática de relações usando métodos de AM supervisionados. Os resultados dos experimentos mostrados na Seção 4.2 mostraram que abordagens baseadas em AM são uma direção promissora para estudos com a língua portuguesa em foco.

A ARS é uma ferramenta de código aberto (licença GPLv3) que pode ser modificada e estendida para acoplar outras relações semânticas, formato de entrada/saída de dados e modelos de AM treinados previamente. Seu código está disponível livremente no portal de tradução automática, PorTAl, em: http://www.lalic.dc.ufscar.br/portal. A ARS não deve causar problemas inesperados, mas caso ocorram, verifique o arquivo ars.log na mesma pasta em que a ferramenta estiver sendo executada para obter maiores informações.

Agradecimentos

Agradecemos ao CNPq e à FAPESP pelo apoio financeiro no desenvolvimento deste trabalho.

6 Referências Bibliográficas

Aziz, W.; Specia, L. (2011). Fully Automatic Compilation of Portuguese-English and Portuguese-Spanish Parallel Corpora. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL-2011)*, Cuiabá, Brasil.

Bick, E. (2000) *The Parsing System"Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Denmark. Aarhus University Press.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. (2009) The weka data mining software: An update. In *SIGKDD Explorations*, v. 11, n. 1, p. 10–18, June 2009.

Ide, N.; Bonhomme, P.; Romary, L. (2000) Xces: An xml-based encoding standard for linguistic corpora. In *Proceedings of LREC 2000*. Atenas, Grécia: [s.n.], p. 825–830.

Joachims, T. (1998) *Making Large-Scale SVM Learning Practical*. [S.l.].

Mengel, A.; Lezius, W. (2000) An xml-based encoding format for syntactically annotated corpora. In *Proceedings of LREC 2000*. Atenas, Grécia: [s.n.], p. 121–126.

Minsky, M. (1986) *The Society of Mind*. [S.l.]: Simon and Schuster.

Quinlan, J. R. (1993) *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kauf-mann Publishers Inc.. ISBN 1-55860-238-0.

Taba, L. S.; Caseli, H. de M. (2012) Uma ferramenta para anotação de relações semânticas entre termos. In *Anais do XI Encontro de Linguística de Corpus – ELC 2012*. Instituto de Ciências Matemáticas e de Computação da USP, São Carlos/SP, Brasil: [s.n.].

Taba, L. S. (2013) *Extração automática de relações semânticas a partir de textos escritos em português do Brasil*, São Carlos/SP, Brasil.

Vapnik, V. (1995) *The nature of statistical learning theory*. [S.l.]: Springer Verlag.