Instituto de Ciências Matemáticas e de Computação – ICMC-USP

Identificação Automática de Microaspectos em Textos Jornalísticos

Alessandro Yovan Bokan Garay
Thiago Alexandre Salgueiro Pardo

NILC-TR-15-01

Abril, 2015

Série de Relatórios do
Núcleo Interinstitucional de Linguística Computacional (NILC)

NILC-ICMC-USP, Caixa Postal 668, 13560-970, São Carlos, SP, Brasil

Resumo

Os aspectos informativos representam as unidades básicas de informação presentes nos textos. Por exemplo, em textos jornalísticos em que se relata um fato/acontecimento, os aspectos podem representar a seguintes informações: o que aconteceu, onde aconteceu, quando aconteceu, como aconteceu, e por que aconteceu. Com a identificação dos aspectos é possível automatizar algumas tarefas do Processamento da Linguagem Natural (PLN), como Sumarização Automática, Perguntas e Respostas e Extração de Informação. Segundo Rassi et al. (2013), aspectos podem ser de 2 tipos: *microaspectos* e *macroaspectos*. Os *microaspectos* representam os segmentos locais das sentenças. Já os *macroaspectos* emergem da informação contida nas sentenças em seus contextos. Neste relatório, descreve-se a metodologia e os resultados do processo de identificação de *microaspectos* utilizando duas abordagens: uso de sistemas como anotador de papeis semânticos, reconhecedor de entidades mencionadas e regras manuais; uso de técnicas de aprendizado de máquina. A metodologia foi avaliada sobre o córpus de notícias jornalísticas CSTNews, previamente anotado manualmente com aspectos informativos. Os resultados são satisfatórios e demostram que os *microaspectos* podem ser identificados automaticamente em textos jornalísticos com um desempenho razoável.

Conteúdo

1. Introdução

Este relatório apresenta o processo de identificação automática de "aspectos informativos" em textos jornalísticos. Os aspectos representam componentes semântico-discursivos que correspondem às unidades básicas de informação presentes nas sentenças dos textos do gênero jornalístico. Os aspectos podem representar componentes locais da sentença, indicando informações, tais como local específico ou uma data determinada; também podem ser frutos das relações discursivas entre os segmentos de um texto. Em uma notícia jornalística sobre desastres naturais, por exemplo, os seguintes aspectos poderiam ser identificados: "quando aconteceu", "onde aconteceu", "o que aconteceu", "quais foram as contramedidas".

Os aspectos surgiram no âmbito da *Text Analysis Conference*[1] (TAC), a principal conferência e competição científica dedicada à Sumarização Automática (SA). Nessa conferência, Owczarzak e Dang (2011) propuseram a utilização de "aspectos informativos" como uma abordagem profunda para a produção de sumários multidocumento. Segundo os autores, os aspectos podem ser úteis para a elaboração de sumários coerentes e direcionados para o gênero e categoria textual em foco. No total, foram definidas cinco categorias: "Acidentes e desastres Naturais", "Ataques", "Saúde e Segurança", "Recursos em via de extinção" e "Julgamentos e investigações". Assim, as categorias indicam o assunto ou domínio do texto. Como ilustração, a TAC propõe que os sumários da categoria "Ataques" contenham os aspectos WHAT, WHEN, WHERE, WHY, WHO_AFFECTED, DAMAGES, PERPETRATORS e COUNTERMEASURES[2]. Como exemplo, na Fig. 1, apresenta-se um sumário multidocumento da categoria "Ataques", anotado manualmente com aspectos informativos. Na primeira sentença do sumário, informa-se que uma série de ataques criminosos (WHAT) aconteceram na cidade de São Paulo (WHERE) na segunda-feira, 7 (WHEN). Na segunda sentença, identificam-se as entidades afetadas pelos ataques (WHO_AFFECTED). Já na última sentença, identificam-se as entidades criminosas (PERPETRATORS).

[Uma nova série de ataques criminosos foi registrada na madrugada desta segunda-feira, dia 7, em São Paulo e municípios do interior paulista.] WHAT/WHEN/WHERE

[Os bandidos atacaram agências bancárias, bases policiais e prédios públicos com bombas e tiros.] WHO_AFFECTED

[As ações são atribuídas à facção criminosa Primeiro Comando da Capital (PCC), que já comandou outros ataques em duas ocasiões.] PERPETRATORS

Figura 1: Sumário da categoria "Ataques" anotado com aspectos

Genest et al. (2009) afirmam que a identificação de aspectos pode ser útil tanto para a determinação de informações relevantes dos textos-fonte quanto para a identificação de restrições estruturais na construção dos sumários. A partir de sua adoção na TAC, os aspectos foram utilizados em vários trabalhos da literatura para auxiliar a tarefa de sumarização (Steinberger et al., 2010; Li et al., 2011; Genest e Lapalme, 2012). Porém, o uso de aspectos não é novidade em sumarização e nem em outras áreas do Processamento de Linguagem Natural (PLN). Por exemplo, Swales (1999) propõe o uso de aspectos como componentes

[1] http://www.nist.gov/tac

[2] Terminologia em inglês proposta pela TAC.

semânticos e discursivos aplicados no modelo CARS (*Create a Research Space*) na forma de estruturas esquemáticas para construir/estruturar textos científicos. Alguns trabalhos pioneiros em sumarização que usaram o conceito de aspectos informativos são os trabalhos de Teufel e Moens (1999, 2002) e White et al. (2001). Acredita-se também que os aspectos possam auxiliar outras tarefas relacionadas, como Mineração de texto, por exemplo.

Com base nos aspectos informativos, identificam-se estruturas de seleção/organização de conteúdo para a construção de sumários, sendo possível gerar sumários de qualidade com informações de interesse para o usuário final. Portanto, neste trabalho, a finalidade de se identificar aspectos automaticamente é de **auxiliar** no processo de geração automática de sumários com base nas estruturas previamente definidas (Rassi et al., 2013). Cabe ressaltar que este trabalho é parte do processo de sumarização multidocumento de um projeto de mestrado.

Como já foi dito, os aspectos são definidos conforme os diferentes gêneros textuais: jornalístico, opinião, científico, literário, etc. Neste trabalho, os aspectos informativos estão definidos especificamente para o **gênero jornalístico**, com base na tarefa de Sumarização promovida pela TAC. Relata-se, então, o processo e os resultados obtidos na identificação automática de aspectos informativos no córpus de notícias jornalísticas CSTNews (Cardoso et al., 2011). O restante do trabalho está organizado da seguinte forma: na Seção 2, apresentam-se os recursos e sistemas que serão utilizados no processo de identificação; na Seção 3, descreve-se a metodologia utilizada; na Seção 4, mostram-se os resultados obtidos na identificação; e por fim, na Seção 5, apresentam-se as conclusões.

2. Recursos e sistemas

2.1. O córpus CSTNews

O córpus CSTNews[3] (Cardoso et al., 2011) é um recurso composto por coleções de textos-fonte de gênero jornalístico, construído com vistas à investigação da SA mono e multidocumento para o português brasileiro. O córpus contém 50 coleções de textos jornalísticos. Cada coleção engloba de 2 a 3 textos sobre um mesmo assunto. Os textos foram compilados manualmente dos jornais *online* "Folha de São Paulo", "O Globo", "Jornal do Brasil", "Estadão" e "Gazeta do Povo". As coleções foram classificadas em 6 categorias textuais: *Cotidiano*, *Esporte*, *Mundo*, *Política*, *Dinheiro* e *Ciência*. Cada categoria contém uma determinada quantidade de coleções de textos jornalísticos (ver Fig. 2). Assim, foram identificadas 14 coleções da categoria *Cotidiano*, 10 coleções da categoria *Esporte*, 14 coleções da categoria *Mundo*, 10 coleções da categoria *Política*, 1 coleção da categoria *Dinheiro* e 1 coleção da categoria *Ciência*.

[3] http://www.icmc.usp.br/pessoas/taspardo/sucinto/cstnews.html

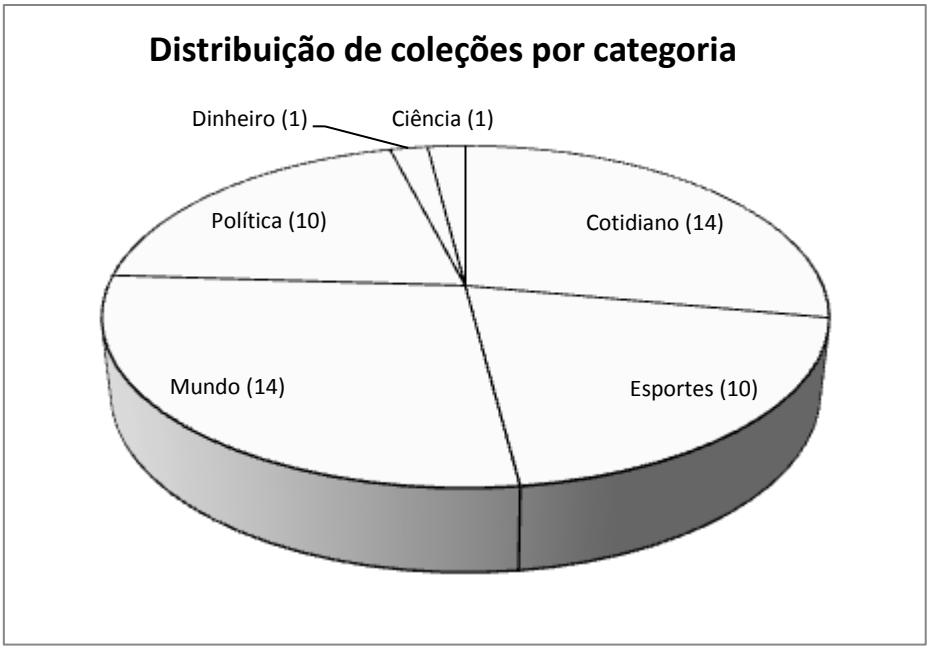

Figura 2: Distribuição das coleções para cada categoria

Além dos textos-fonte crus, o córpus CSTNews possui um total de 140 sumários manuais abstrativos monodocumento, 50 sumários manuais abstrativos multidocumento e 50 sumários manuais extrativos multidocumento. Também existem versões anotadas, em nível discursivo, dos textos-fonte com base na *Rhetorical Structure Theory* (RST) (Mann e Thompson, 1987) e na *Cross-document Structure Theory* (CST) (Radev, 2000), além de várias outras anotações. Na subseção seguinte, descreve-se a anotação manual de aspectos informativos sobre os 50 sumários manuais multidocumento do córpus CSTNews.

2.1.1. CSTNews: aspectos

A tarefa de anotação de córpus é uma tarefa de classificação que consiste em atribuir um ou mais rótulos a uma unidade representativa do texto (palavra, sentença ou parágrafo, normalmente). A anotação de aspectos informativos foi feita por Rassi et al. (2013) em **nível sentencial** sobre os sumários manuais multidocumento do córpus CSTNews. Para a tarefa de SA multidocumento, os aspectos podem indicar estruturas padrão para a modelagem de critérios de seleção e organização de conteúdo nos sumários. As categorias no córpus CSTNews diferem das definidas na TAC 2010. Contudo, existem similaridades com as seis categorias consideradas (ver Fig. 2). Por exemplo, nas categorias *Cotidiano* ou *Mundo*, pode haver menção a "*Acidentes e desastres naturais*".

A tarefa de anotação foi realizada por 4 subgrupos de anotadores composto por 3 ou 4 linguistas computacionais, havendo um pesquisador sênior em cada subgrupo responsável pela coordenação da tarefa de anotação. Cada subgrupo ficou responsável pela anotação completa de uma das 4 categorias mais representativas, ou com maior quantidade de textos-fonte do córpus (*Cotidiano*, *Esportes*, *Mundo*, *Política*). Na **fase preliminar** de anotação, para ter uma referência consensual, foram anotados os sumários das categorias *Dinheiro* (1) e *Ciência* (1). Já na **fase final** de anotação, foram anotados os 48 sumários das categorias *Cotidiano* (14), *Esporte* (10), *Mundo* (14) e *Política* (10).

Com base na tarefa de anotação definida pela TAC, aplicou-se um refinamento e definição dos aspectos em função das diferentes categorias sugeridas nos textos-fonte. Esse refinamento envolveu tanto a exclusão de algumas etiquetas originais quanto a inserção de novas etiquetas de interesse para os textos do córpus CSTNews. Assim, foram definidos 20 aspectos informativos que podem se referir a conceitos ou objetos[4] (ver Tab. 1).

Macroaspectos	Microaspectos
COMMENT	WHO_AGENT
COMPARISON	WHO_AFFECTED
CONSEQUENCE	WHEN
COUNTERMEASURES	WHERE
DECLARATION	WHY
GOAL	HOW
HISTORY	SCORE
PREDICTION	SITUATION
SITUATION	GOAL
WHAT	
HOW	

Tabela 1: Aspectos gerais do córpus CSTNews

A necessidade de identificação de segmentos textuais em diversos níveis estruturais para a determinação do aspecto correspondente resultou na classificação dos aspectos em *microaspectos* e *macroaspectos*. Os *microaspectos* representam segmentos locais que compõem uma sentença (segmentos *intrassentenciais*). Os *macroaspectos* dependem do conteúdo sentencial em contexto. No total, foram identificados 11 *macroaspectos* e 9 *microaspectos*, apesar de haver alguma variação nesses conjuntos, como discutimos a seguir. É importante dizer que os aspectos foram analisados e classificados de maneira diferente para cada categoria em particular. Assim, foram excluídos os aspectos que não ocorrem como *microaspectos* ou como *macroaspectos* nos sumários anotados (ver Apêndice A). Por exemplo, a categoria *Esportes* é a única que possui o *microaspecto* SCORE. Nota-se que os aspectos SITUATION, GOAL e WHO podem acontecer tanto como *macroaspectos* quanto como *microaspectos*.

Cabe ressaltar que a anotação de aspectos foi feita em **nível sentencial**, seguindo a metodologia da TAC, ou seja, os aspectos identificados são posicionados ao final da sentença. Na Fig. 3, mostra-se um exemplo de uma sentença anotada com aspectos da categoria *Mundo*. Com respeito aos *macroaspectos*, descreve-se o acontecimento de um desastre natural (WHAT) e a declaração emitida pelo jornal japonês pró-Pyongyang (DECLARATION). Com respeito aos *microaspectos*, informa-se que o fato aconteceu no mês de julho (WHEN), na Coréia no Norte (WHERE), por causa das enchentes (WHY), deixando muitas pessoas mortas e outras feridas (WHO AFFECTED).

Por último, na anotação foi relevante distinguir aspectos que transmitam informações principais daqueles relativos a informações secundárias. Diante disso, os aspectos foram referenciados pelo sufixo EXTRA. Por exemplo, se uma sentença possui alguma informação de local, mas que não se refere ao evento principal, então é anotada como WHERE_EXTRA. Neste

[4] Para efeito de divulgação, optou-se por manter a terminologia em inglês da TAC.

trabalho, não existe uma distinção entre ideias principais e secundarias. Portanto, os sufixos EXTRA foram ignorados, deixando o aspecto na sua forma original (ver Tab. 1).

[Ao menos 549 pessoas morreram, 3.043 ficaram feridas e outras 295 ainda estão desaparecidas em consequência das enchentes que atingiram a Coréia do Norte em julho, segundo um jornal japonês pró-Pyongyang.] **WHO_AFFECTED/WHAT/WHY/WHERE/WHEN/ DECLARATION**

Figura 3: Sentença anotada do sumário da coleção C1 do córpus CSTNews

2.2. O *parser* PALAVRAS

O *parser* PALAVRAS é um analisador sintático de textos em língua portuguesa baseado em regras, desenvolvido por Bick (2000). O PALAVRAS segue a metodologia da Gramática de Constituintes (em inglês, *Constraint Grammar*) introduzido por Karlsson (1990), a fim de resolver problemas de ambiguidade morfológica e mapear funções sintáticas por meio da dependência de contexto.

O *parser* pode transformar uma notação de Gramática de Constituintes (formato *flat*) em uma estrutura de árvore sintática tradicional (formato *tree*). Na Fig. 4, ilustra-se um exemplo de anotação simples da sentença "O menino nada na piscina". Dentro dos colchetes ([]), encontra-se a palavra na forma lematizada. Em seguida, aparecem os rótulos semânticos entre os símbolos "<" e ">". Logo depois, são anotadas as classes gramaticais, como substantivo (N), verbo (V), determinante (DET) e preposição (PREP). Junto com as classes gramaticais, estão as informações morfossintáticas indicando, por exemplo, que o verbo "nadar" está no tempo presente (PR), na terceira pessoa do singular (3S), do modo indicativo (IND), flexionado (VFIN). Por último, após o símbolo "@", indicam-se as funções sintáticas. Por exemplo, a palavra "menino" foi marcada com @SUBJ, que significa o sujeito da oração.

```
O [o] <artd> DET M S @>N
menino [menino] <H> N M S @SUBJ>
nada [nadar] <fmc> <mv> V PR 3S IND VFIN @FS-STA
em [em] <sam-> PRP @<ADVL
a [o] <artd> <-sam> DET F S @>N
piscina [piscina] <Lh> N F S @P<
\$.
```

Figura 4: Anotação de Gramática de Constituintes simples (*flat*)

Segundo seu autor, usando um conjunto de etiquetas gramaticais bastante diversificado, o *parser* alcança um nível de correção (ou exatidão) de 99% em termos de morfossintaxe (classe gramatical e flexão), e 97-98% em termos de sintaxe. Na prática, tem se verificado desempenho inferior a esse relatado. Neste trabalho, o *parser* PALAVRAS foi utilizado tanto como entrada do classificador de papéis semânticos (ver Seção 3.1 e Seção 3.2) quanto para fornecer informações morfossintáticas e semânticas na criação do classificador de aspectos usando métodos de Aprendizado de Máquina (ver Seção 3.4).

2.3. Anotação de papeis semânticos

Para compreender a semântica de uma sentença, deve-se analisar o comportamento do verbo em relação aos constituintes (ou argumentos) que a envolvem. Assim, chamam-se de "papéis semânticos" as relações semânticas entre o verbo e seus argumentos. A tarefa de identificar quais grupos de palavras atuam como argumentos de um determinado verbo é chamada de "Anotação de Papéis Semânticos (APS)" (Shamsfard e Mousavi, 2008).

Para o português brasileiro, existe o trabalho feito por Alva-Manchego (2013). O autor propõe um sistema de classificação que consta de 3 fases: (1) identificação do verbo alvo, (2) identificação de argumentos e (3) classificação de argumentos. Na Fig. 5, ilustra-se um exemplo do processo de anotação de papéis semânticos. Em primeiro lugar, identifica-se o verbo alvo "venceu" (*v*). Em seguida, identificam-se os argumentos (*A*) "a equipe brasileira", "a Finlândia" e "em Tampere". Por último, os argumentos são anotados com os papéis semânticos "A0" (sujeito agente da oração), "A1" (sujeito paciente da oração) e "AM-LOC" (local da ação), respectivamente. Cabe ressaltar que a terminologia "A/Arg" refere-se ao "argumento" identificado, seguido de um numero prototípico. Já a terminologia "AM/ArgM" refere-se ao "argumento modificador", seguido do tipo de modificador, como tempo, local, maneira, causa, etc.

A equipe brasileira [*venceu* $_{v}$] a Finlândia em Tampere. (1)
[A equipe brasileira $_{A}$] [*venceu* $_{v}$] [a Finlândia $_{A}$] [em Tampere $_{A}$]. (2)
[A equipe brasileira $_{A0}$] [*venceu*] [a Finlândia $_{A1}$] [em Tampere $_{AM\text{-}LOC}$]. (3)

Figura 5: Exemplo de anotação de papéis semânticos

O resultado final do sistema, conforme a medida F1 (ver Seção 4), foi de 94.5% na fase de identificação de argumentos, 81.70% na fase de classificação de papéis semânticos e 79.70% na fase de identificação + classificação.

Os constituintes ou argumentos relacionados ao verbo podem responder a perguntas do tipo quem?, quando?, onde? e como?. No exemplo anterior, a resposta à pergunta "quem venceu?" seria "A equipe brasileira". De maneira igual, as perguntas "quem foi vencido?" e "onde foi vencido?" seriam respondidas por "a Finlândia" e "em Tampere", respetivamente. Tais constituintes podem definir aspectos informativos no nível *microsentencial*, como WHO_AGENT ("quem venceu"), WHO_AFFECTED ("quem foi vencido") e WHERE ("onde"), respectivamente. Assim, os papéis semânticos são, normalmente, similares aos *microaspectos*. Dessa maneira, neste trabalho de pesquisa, foi proposto o uso do APS para o português brasileiro feito por Alva-Manchego (2013). Na Tab. 2, apresentam-se as equivalências propostas entre alguns *microaspectos* e os papéis semânticos usados por Alva-Manchego (2013) e definidos por Palmer et al. (2010).

Microaspecto	Papel semântico	Nome	Definição
WHO_AGENT	A0 / Arg2	Agente	O sujeito da ação
WHO_AFFECTED	A1 / Arg1	Paciente	O afetado pela ação
WHERE	AM-LOC / ArgM-LOC	Local	Onde ocorreu a ação
WHEN	AM-TMP / ArgM-TMP	Tempo	Quando ocorreu a ação
HOW	AM-MNR / ArgM-MNR	Maneira	Como a ação foi realizada
WHY	AM-CAU / ArgM-CAU	Causa	Causa ou motivo da ação

Tabela 2: Equivalências entre *microaspectos* e papéis semânticos

2.4. Reconhecimento de entidades mencionadas

Em geral, as entidades mencionadas (EM) são entidades concretas ou abstratas referenciadas no texto por um nome próprio. Outros elementos também são considerados como EMs, como o caso das datas, por exemplo. O termo EM foi cunhado pela "*Sixth Message Understanding Conference* (MUC-6)" (Grishman e Sundheim, 1996). O Reconhecimento de Entidades Mencionadas (REM) é uma subtarefa da Extração da Informação (EI) que visa identificar e classificar entidades do texto em categorias pré-definidas, tais como PESSOA, ORGANIZAÇÃO, LOCAL, TEMPO, VALOR e ACONTECIMENTO, entre outras categorias de interesse (Nadeau e Sekine, 2007).

Nesse contexto, é importante citar o HAREM[5] (Santos e Cardoso, 2007), que é um evento de avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para coleções de documentos em português organizado pela Linguateca[6]. No âmbito do primeiro HAREM[7], vários trabalhos foram apresentados, sendo o PALAVRAS_NER (Bick, 2007) o sistema que obteve os melhores resultados nas tarefas de identificação e classificação de EM. O PALAVRAS_NER baseia-se no analisador morfossintático PALAVRAS (ver Seção 2.2) para criar um conjunto de regras manuais que identificam EM nos textos. Já no contexto do segundo HAREM[8], um dos sistemas "*open source*" com os melhores resultados foi o REMBRANDT (Cardoso, 2008).

O REMBRANDT[9] (**R**econhecimento de **E**ntidades **M**encionadas **B**aseado em **R**elações e **AN**álise **D**etalhada do **T**exto) é um sistema de REM e de Detecção de Relações entre Entidades (DRE) para o português. Segundo Cardoso (2008), o REMBRANDT explora intensamente a Wikipédia como fonte de conhecimento, e aplica um conjunto de regras gramaticais que aproveitam os vários indícios internos e externos das EM para extrair o seu significado. Além disso, o REMBRANDT possui uma interface própria para interagir com a Wikipédia, a SASKIA, com o objetivo de facilitar as tarefas de navegação na estrutura de categorias, ligações e redirecionamentos da Wikipédia, com vista à extração de conhecimento. No contexto do segundo HAREM, o sistema REMBRANDT teve um desempenho de 56.74% de medida F1 na fase de REM e 45.02% na fase de DRE.

Na Fig. 6, ilustra-se uma sentença anotada pelo sistema REMBRANDT. Para efeito de visualização, as EM foram anotadas com as etiquetas "<EM></EM>", indicando a categoria (C) a qual a entidade pertence. Nessa sentença, a entidade "Bernardinho" foi identificada como PESSOA, a entidade "Finlândia" como LOCAL, as entidades "3" e "0" como NÚMERO e a entidade "Jogos Pan-Americanos" como ACONTECIMENTO. Nota-se que a entidade "Finlândia" foi anotada erroneamente como LOCAL porque, no contexto, faz referência a uma equipe de vôlei.

[5] **H**AREM refere-se à **A**valiação de **R**econhecedores de **E**ntidades **M**encionadas.
[6] http://www.linguateca.pt/
[7] http://www.linguateca.pt/primeiroHAREM/harem.html
[8] http://www.linguateca.pt/harem/
[9] http://xldb.di.fc.ul.pt/Rembrandt/

A equipe brasileira, comandada por <EM C="PESSOA"> Bernardinho </EM>, venceu a <EM C="LOCAL"> Finlândia </EM> por <EM C="NÚMERO">3</EM> sets a <EM C="NÚMERO"> 0 </EM>, nos <EM C="ACONTECIMENTO"> Jogos Pan-Americanos </EM>.

Figura 6: Exemplo de anotação do sistema REMBRANDT

Da mesma forma que os papéis semânticos, as categorias das entidades mencionadas podem definir alguns aspectos informativos no nível *microsentencial*: o aspecto WHERE é equivalente a LOCAL, WHEN a TEMPO e SITUATION a ACONTECIMENTO. Dessa maneira, neste trabalho de pesquisa, foi proposto o uso do sistema REMBRANDT na identificação automática de alguns *microaspectos*. Na Tab. 3, apresentam-se as equivalências propostas entre os *microaspectos* e as categorias das EM.

Microaspecto	Categoria EM
WHERE	LOCAL
WHEN	TEMPO
SITUATION	ACONTECIMENTO

Tabela 3: Equivalências entre microaspectos e entidades nomeadas

3. Metodologia

Neste relatório, o processo de identificação automática de *microaspectos* foi dividido em 3 fases (ver Fig. 7). A seguir, explicam-se as fases do processo de identificação:

1. Compilar as sentenças dos 48 sumários anotados do córpus CSTNews das categorias *Cotidiano*, *Esportes*, *Mundo* e *Política*. Não foram consideradas as categorias *Dinheiro* e *Ciência*, por terem poucos textos anotados.
2. Anotar automaticamente as sentenças com *microaspectos* usando 4 sistemas diferentes:
    a. **Sistema APS (Anotador de Papéis Semânticos):** sistema que anota automaticamente sentenças com os *microaspectos* equivalentes aos papéis semânticos apresentados na Tab. 2. Abrangem-se os *microaspectos* WHO_AGENT, WHO_AFFECTED, WHEN, WHERE, WHY e HOW.
    b. **Sistema APS + Regras:** sistema que usa regras desenvolvidas manualmente neste trabalho com base nos *"falsos negativos e positivos"* do sistema APS, com a finalidade de aprimorar o seu desempenho. Abrangem-se os *microaspectos* WHO_AGENT, WHO_AFFECTED, WHEN, WHERE, WHY e SCORE.
    c. **Sistema REMBRANDT:** sistema que anota automaticamente sentenças com os *microaspectos* equivalentes às categorias das entidades mencionadas apresentadas na Tab. 3. Abrangem-se os *microaspectos* WHEN, WHERE e SITUATION.
    d. **Aprendizado de Máquina (AM):** uso de técnicas de AM para criar um classificador de *microaspectos*. Atende todos os *microaspectos*, com exceção do aspecto GOAL.
3. Obter um conjunto de sentenças anotadas automaticamente com *microaspectos*.

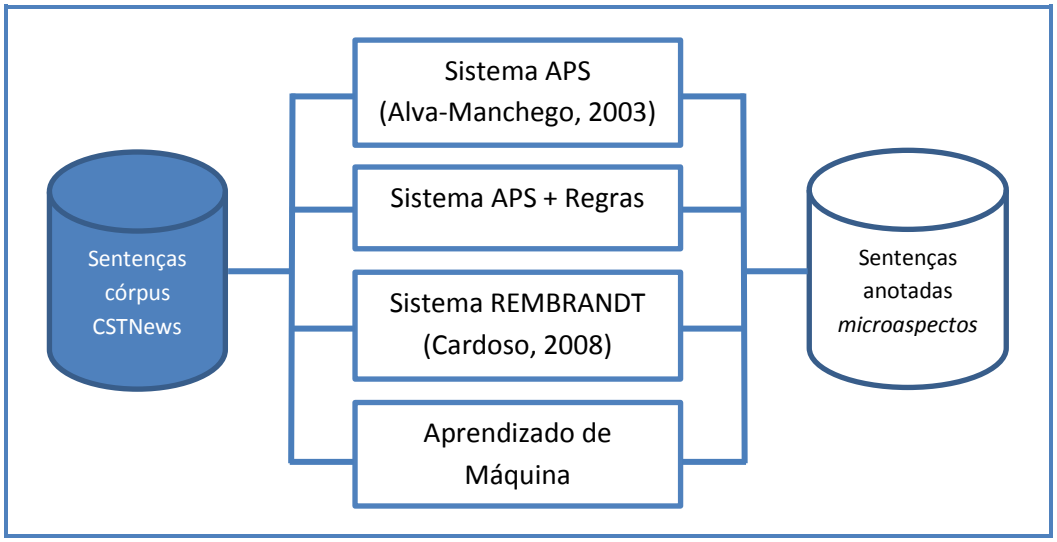

Figura 7: Metodologia do processo de identificação de microaspectos

Cabe ressaltar que o *microaspecto* GOAL **não foi considerado**, por não ser identificado por nenhum dos sistemas utilizados: APS, APS + Regras, REMBRANDT e AM. A seguir, descrevem-se detalhadamente os sistemas de anotação automática de *microaspectos*.

3.1. O sistema APS

O sistema APS (Anotador de Papéis Semânticos) é composto por 2 subsistemas principais: o *parser* PALAVRAS (Bick, 2000) e o classificador de papéis semânticos (Alva-Manchego, 2013). Basicamente, o parser PALAVRAS gera as árvores sintáticas que servem de instâncias para que o classificador possa anotar as sentenças com os papéis semânticos correspondentes. Já no final, os papéis semânticos são mapeados nos *microaspectos* WHO_AGENT, WHO_AFFECTED, WHEN, WHERE, WHY e HOW, conforme apresentado na Tab. 2. A seguir, relata-se o processo efetuado pelo sistema APS:

1. Dado um conjunto de sentenças a serem anotadas, utiliza-se o *parser* PALAVRAS a fim de gerar árvores sintáticas para cada sentença. Tais árvores são representadas em formato *TigerXML*[10] (ou *tree*). Neste passo, o PALAVRAS vai pré-processar as sentenças de entrada. Assim, pode se dar o caso de separação das preposições, por exemplo: "dos" em "de os", "nesta" em "em esta", "pelo" em "por o", etc.
2. Executa-se um algoritmo que clona as árvores sintáticas conforme o número de verbos alvo da sentença (sem considerar os verbos auxiliares). Desta maneira, se uma sentença possui três verbos alvo, a árvore sintática da sentença será clonada duas vezes, indicando o verbo alvo correspondente para cada árvore (ou instância).
3. Classifica-se cada árvore/instância da sentença pelo anotador de papéis semânticos de Alva-Manchego (2013).
4. Mapeiam-se os papéis semânticos nos *microaspectos* correspondentes.
5. Executa-se um programa que transforma o formato de saída convencional do classificador, denominado formato CoNLL[11], em um formato de rótulos "<aspect>

[10] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/doc/html/TigerXML.html
[11] http://ilk.uvt.nl/conll/

</aspect>". Assim, os *microaspectos* são anotados na sentença em um formato mais legível para o usuário.

6. Posiciona-se, por fim, os aspectos anotados no final da sentença.

Para exemplificar a saída do sistema APS, na Fig. 8, ilustra-se uma sentença anotada automaticamente com *microaspectos*. Nota-se que o segmento "A equipe brasileira" foi anotado como WHO_AGENT por representar o sujeito gramatical agente da oração e por estar relacionado semanticamente ao verbo "vencer (venceu)". Da mesma forma, os segmentos "a Finlândia" e "em Tampere" foram anotados como WHO_AFFECTED (paciente) e WHERE (local), respectivamente.

<aspect APS="WHO_AGENT">A equipe brasileira</aspect> **venceu** <aspect APS="WHO_AFFECTED">a Finlândia</aspect> <aspect APS="WHERE">em Tampere</aspect>.

Figura 8: Sentença anotada com *microaspectos* pelo sistema APS

Por último, a anotação de *microaspectos* será feita em nível sentencial (ver Fig. 9). Isso significa que não importa se um *microaspecto* é anotado várias vezes numa mesma sentença (presença de vários verbos alvo), pois ele será anotado uma vez só no final da sentença.

[A equipe brasileira venceu a Finlândia em Tampere.]**WHO_AGENT, WHO_AFFECTED,WHERE**

Figura 9: Anotação de *microaspectos* em nível sentencial usando o sistema APS

3.2. O sistema APS + Regras

Tal sistema consiste de um conjunto de regras criadas manualmente com base nos padrões presentes nas sentenças identificadas como "falsos negativos" e "falsos positivos" da anotação feita pelo sistema APS. Os primeiros referem-se às sentenças cujos aspectos o sistema APS não conseguiu classificar, mas que foram anotadas manualmente. Já os segundos referem-se às sentenças que o sistema APS conseguiu classificar, mas que não foram anotadas manualmente. Assim, foram criadas regras para os aspectos WHO_AGENT, WHO_AFFECTED, WHEN, WHERE e WHY. Porém, não foram criadas regras para os aspectos HOW, SITUATION e GOAL, já que não foram encontrados padrões que possam identificar esses aspectos. Já para o aspecto SCORE, também foram criadas regras manuais, mesmo não tendo equivalência com algum papel semântico. No Apêndice B, apresenta-se o conjunto de regras.

Tanto as regras do aspecto WHO_AGENT quanto as regras do aspecto WHO_AFFECTED estão baseadas somente nos "falsos positivos", pois tais *microaspectos* representam a entidade pessoa ou organização. Porém, o classificador de papéis semânticos não distingue se o agente/paciente da oração é representado por uma pessoa ou organização. Assim, existe uma grande possibilidade do classificador gerar muitos "falsos positivos". Para solucionar este problema, foram usadas as "etiquetas semânticas para substantivos" fornecidas pelo PALAVRAS[12]. A ideia é apagar todos os segmentos identificados como *microaspectos* que não

[12] http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html#semtags_nouns

representem uma entidade (pessoa ou organização) e, assim, diminuir a quantidade de "falsos positivos" (ver Fig. 20).

Para os aspectos WHEN, WHERE e WHY, foram criadas regras que possam identificar automaticamente expressões de tempo, local e causa (respectivamente). As regras do aspecto WHEN foram feitas com base nas regras propostas por Baptista et al. (2008) (ver Fig. 21). Já as regras dos aspectos WHERE (ver Fig. 22) e WHY (ver Fig. 23) foram criadas com o auxílio da pesquisadora colaboradora Magali Sanches Duran[13].

Por último, as regras do aspecto SCORE foram criadas com base nos padrões identificados nas únicas 10 sentenças anotadas no córpus CSTNews (ver Fig. 24). Tais regras foram integradas no sistema APS + Regras.

3.3. O sistema REMBRANDT

O sistema REMBRANDT, feito por Cardoso (2008), visa identificar automaticamente as entidades mencionadas presentes nos textos-fonte. Neste trabalho, o sistema REMBRANDT será utilizado na identificação dos *microaspectos* WHEN, WHERE e SITUATION, por serem equivalentes às entidades TEMPO, LOCAL e ACONTECIMENTO, respectivamente (ver Tab. 3).

Na Fig. 10, ilustra-se um exemplo de sentença anotada com entidades mencionadas. Observa-se que a entidade "Jogos Pan-Americanos" foi reconhecida como ACONTECIMENTO, a entidade "terça-feira" como TEMPO e as entidades "Finlândia" e "Maracanãzinho", como LOCAL. Devido à relação das entidades com os *microaspectos*, os rótulos inseridos pelo sistema REMBRANDT (ver Fig. 6), foram trocados pelos rótulos "<aspect></aspect>".

> No contexto dos <aspect EM="ACONTECIMENTO"> Jogos Pan-Americanos </aspect>, a equipe brasileira de vôlei venceu nesta <aspect EM="TEMPO"> terça-feira</aspect> a <aspect EM="LOCAL">Finlândia</aspect> por 3 sets a 0 no, <aspect EM="LOCAL"> Maracanãzinho</aspect>.

Figura 10: Sentença anotada com entidades mencionadas pelo sistema REMBRANDT

Da mesma maneira que o sistema APS, a anotação será feita no nível sentencial, com as entidades mencionadas mapeadas nos *microaspectos* (ver Fig. 11).

> [No contexto dos Jogos Pan-Americanos, a equipe brasileira de vôlei venceu nesta terça-feira a Finlândia por 3 sets a 0, no Maracanãzinho.] **SITUATION/WHEN/WHERE**

Figura 11: Anotação de *microaspectos* em nível de sentença usando REMBRANDT

[13] http://buscatextual.cnpq.br/buscatextual/visualizacv.do?metodo=apresentar&id=K4737951E6

3.4. Aprendizado de Máquina (AM)

Na atualidade, destaca-se a capacidade dos computadores de aprender tarefas automaticamente com base em alguma experiência. Essa experiência se constrói por meio de um conjunto de exemplos denominados instâncias. Cada instância contém certos *atributos* que representam o conhecimento da tarefa a ser automatizada. Em um sistema de Aprendizado de Máquina (AM), a experiência recebe o nome de *conjunto de treinamento*. Segundo Zhu e Goldberg (2009), a predição desejada em uma instância recebe o nome de rótulo, podendo tornar-se um conjunto finito de valores, denominados *classes*. Em outras palavras, o AM tenta generalizar a predição de uma classe a partir de um conjunto finito de treinamento para dados de teste nunca antes vistos.

Neste trabalho, a tarefa a ser aprendida é a "identificação de *microaspectos*". Devido à disponibilidade de um córpus anotado manualmente (CSTNews), a nossa tarefa segue na linha do paradigma de AM *supervisionado*, em que o conjunto de treinamento está formado por pares *instância-classe* denominados *dados rotulados*. As *instâncias-classes* são as sentenças do córpus anotadas com os aspectos informativos.

No processo, criaram-se vários classificadores com diferentes atributos (em inglês, *features*). Cada atributo foi representado em unigramas (1-grama), bigramas (2-grama) e trigramas (3-grama). Um dos atributos é o conjunto de todas as palavras do córpus (*bag-of-words*). Outros atributos são as informações de *part-of-speech* (POS) e rótulos semânticos (SEM), fornecidos pelo parser PALAVRAS (ver Fig. 4). Os classificadores criados foram o resultado das combinações desses atributos. Por exemplo, criou-se um classificador de atributos de tipo "(*2, 2) POS+SEM"* que significa "bigramas de classes gramaticais e rótulos semânticos", ou "*(1, 2) bag_of_words*" que significa "unigramas e bigramas de palavras". Foi considerada uma frequência de ocorrência maior ou igual a 2 para se filtrar os atributos.

A técnica de aprendizado supervisionado utilizada foi o SVM (*Support Vector Machine*) (Vapnik, 2000). A técnica se baseia no princípio de minimização do risco estrutural trabalhando sobre o conceito de *margem*. O SVM realiza a classificação de dados por meio da construção de vários hiperplanos. O termo margem refere-se à distância mínima a partir do hiperplano de separação até as instâncias de dados mais próximas. A técnica visa criar a maior distância possível entre os hiperplanos de separação e as instâncias próximas a eles. O fato de considerar apenas instâncias próximas às margens é uma característica particular da técnica, daí o nome "vetores de suporte". Escolheu-se SVM porque é a técnica mais utilizada na literatura para classificação com textos. Além disso, SVM é a melhor técnica em tratamento de vetores especiais de grandes dimensões.

Portanto, propõe-se o uso de AM para criar um classificador que possa identificar automaticamente *microaspectos*. A ideia é avaliar todos os possíveis classificadores gerados a partir da combinação dos atributos propostos, usando a técnica SVM, com a finalidade de obter o melhor classificador para cada *microaspecto* em particular. As instâncias de treino e teste são as sentenças dos sumários anotados do córpus CSTNews.

Para finalizar esta seção, na Tab. 4, mostram-se os sistemas utilizados para cada *microaspecto* em particular. Observa-se que os aspectos WHEN e WHERE foram identificados pelos 4 sistemas. Na seguinte seção, relatam-se os resultados dos sistemas avaliados (ver Seção 4).

Microaspecto	APS	APS + Regras	REMBRANDT	AM
WHO_AGENT	X	X		X
WHO_AFFECTED	X	X		X
WHEN	X	X	X	X
WHERE	X	X	X	X
WHY	X	X		X
HOW	X			X
SITUATION		X	X	X
SCORE		X		X

Tabela 4: Sistemas utilizados para cada *microaspecto*

4. Experimentos e resultados

Os sistemas APS, APS + Regras e REMBRANDT foram testados sobre um conjunto de sentenças anotadas manualmente com aspectos informativos. Tais sentenças foram extraídas dos sumários multidocumento do córpus CSTNews (ver Seção 2.1.1). No total, foram anotadas 322 sentenças nas quatro categorias principais: *Cotidiano* (102), *Esportes* (60), *Mundo* (94) e *Política* (66).

Os resultados serão calculados conforme a matriz de confusão apresentada na Tab. 5. Observa-se que na linha superior da matriz estão as classes preditas (P) pelo sistema. Já na coluna da esquerda estão as classes anotadas manualmente, reais (R). Para ter uma estimativa de erro de classificação, dentro da matriz acham-se as seguintes quantidades:

- **Verdadeiros positivos (VP):** refere-se à quantidade de instâncias que o classificador conseguiu anotar automaticamente e que foram anotadas manualmente.
- **Falsos negativos (FN):** refere-se à quantidade de instâncias que o classificador NÃO conseguiu anotar automaticamente, mas que foram anotadas manualmente.
- **Falsos positivos (FP):** refere-se à quantidade de instâncias que o classificador conseguiu anotar automaticamente, mas que NÃO foram anotadas manualmente.
- **Verdadeiros negativos (VN):** refere-se à quantidade de instâncias em que o classificador NÃO conseguiu anotar automaticamente e que NÃO foram anotadas manualmente.

	Verdadeiro (P)	Falso (P)
Verdadeiro (R)	VP	FN
Falso (R)	FP	VN

Tabela 5: Matriz de confusão

As estimativas de erro são calculadas por meio das quantidades de instâncias/exemplos, dando origem às métricas. As métricas são calculadas conforme as classes positiva (SIM) e negativa (NÃO) para cada microaspecto. A seguir, explicam-se as métricas usadas neste trabalho:

- **Cobertura (classe SIM):** também chamada de "taxa verdadeira positiva". Refere-se à taxa de exemplos verdadeiramente positivos que foram classificados como tal.

$$C = \frac{VP}{VP + FN}$$

- **Cobertura (classe NÃO):** também chamada de "taxa verdadeira negativa" ou "especificidade". Refere-se à taxa de exemplos verdadeiramente negativos que foram classificados como tal.

$$C = \frac{VN}{VN + FP}$$

- **Precisão (classe SIM):** também chamada de "valor preditivo positivo". Refere-se à taxa de exemplos classificados como positivos que efetivamente o são.

$$P = \frac{VP}{VP + FP}$$

- **Precisão (classe NÃO):** também chamada de "valor preditivo negativo". Refere-se à taxa de exemplos classificados como negativos que efetivamente o são.

$$P = \frac{VN}{VN + FN}$$

- **Medida F1:** refere-se à "média harmônica" ponderada da precisão e da cobertura, em que as duas métricas tem o mesmo peso ($\alpha = 1$). O cálculo é feito tanto para a classe positiva quando para a classe negativa.

$$F\alpha = \frac{(1 + \alpha) \times P \times C}{\alpha \times (P + C)}$$

$$F1 = \frac{2 \times P \times C}{P + C}$$

- **Acurácia:** refere-se à taxa do total de acertos (VP + VN) sobre o total de exemplos.

$$P = \frac{VP + VN}{VP + VN + FP + FN}$$

A seguir, apresentam-se os resultados obtidos pelos sistemas propostos para cada *microaspecto* (APS, APS + Regras e REMBRANDT).

4.1. WHO_AGENT

Segundo Rassi et al. (2013), o *microaspecto* WHO_AGENT é definido como *"a entidade (pessoa ou organização) responsável por causar/provocar a ocorrência de um fato/evento".* Portanto, o aspecto WHO_AGENT pode ser representado pelo sujeito gramatical agente da oração, seja pessoa ou organização. Na Fig. 12, ilustra-se um exemplo de uma sentença anotada com o aspecto WHO_AGENT. Nota-se que o agente definido pelo segmento "A equipe brasileira", representado pela entidade organização, está relacionado semanticamente ao verbo "vencer (venceu)".

A equipe brasileira, comandada por Bernardinho, **venceu** a Finlândia por 3 sets a 0, em Tampere (FIN), mantendo sua invencibilidade na Liga Mundial de Vôlei-06.

Figura 12: *Sentença anotada com microaspecto WHO_AGENT*

Para identificar automaticamente o aspecto WHO_AGENT, foram utilizados os sistemas APS e APS + Regras. No entanto, não foi utilizado o sistema REMBRANDT, por ser incapaz de identificar o sujeito agente da oração. O sistema foi testado sobre o córpus CSTNews com um total de 130 sentenças anotadas manualmente com o aspecto WHO_AGENT (ver Tab. 6). Observa-se que o aspecto WHO_AGENT está bem distribuído entre todas as categorias.

Categoria	Frequência
Cotidiano	34
Esportes	32
Mundo	29
Política	35
Total	**130**

Tabela 6: Distribuição do aspecto WHO_AGENT por categoria

Na Tab. 7, apresentam-se os resultados dos sistemas testados. Observa-se que o sistema APS foi um pouco melhor que o sistema APS + Regras. Porém, um grande defeito do sistema APS é anotar qualquer entidade sujeito da oração, mesmo sem saber se aquela entidade é uma pessoa ou organização. Assim, existe a probabilidade do sistema APS ter acertado muito (0.815) e ter errado muito também (0.522). As regras do WHO_AGENT foram criadas com base nos "falsos positivos", com a finalidade de melhorar a precisão. Contudo, essa melhora de 0.664 fez que a cobertura diminuísse a 0.592. Mesmo assim, o sistema APS + Regras é mais confiável, porque consegue saber se o sujeito da ação é uma entidade pessoa/organização, conforme a definição do aspecto WHO_AGENT. Na Tab. 8, apresenta-se a matriz de confusão do sistema APS + Regras. Nota-se que das 130 sentenças anotadas, só 77 foram anotadas corretamente e 53 não foram anotadas. A seguir, analisam-se algumas sentenças identificadas como "falsos negativos" e "falsos positivos" do sistema APS + Regras.

WHO_AGENT	Cobertura	Precisão	F1	Acurácia
APS	0.815	0.522	0.637	0.624
APS + Regras	**0.592**	**0.664**	**0.626**	**0.714**

Tabela 7: Resultados para o *microaspecto* WHO_AGENT

WHO_AGENT	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	77	53	Sim	0.592	0.664	0.626	0.714
Falso	39	153	Não	0.797	0.743	0.769	

Tabela 8: *Matriz de confusão do aspecto WHO_AGENT*

Falsos negativos

Na Tab. 9, mostram-se algumas sentenças cujos aspectos não foram identificados automaticamente pelo sistema APS + Regras, mas que foram anotados manualmente. Pode-se observar, em nível discursivo, que na sentença da categoria *Cotidiano*, a entidade "o presidente" atua como WHO_AGENT, porque foi a entidade que fez a afirmação introduzida

pelo segmento “Segundo”. Contudo, em nível semântico, o sistema não consegue identificar o segmento “o presidente” como agente por não haver um verbo elocutivo (p. ex: “o presidente afirmou que”, “o presidente disse que”). Por outro lado, na categoria *Esportes,* o segmento “o Brasil” representa o agente relacionado ao verbo “estar (está)”, mas, mesmo assim, o classificador não conseguiu identificá-lo como WHO_AGENT. Da mesma maneira, na categoria *Mundo*, o segmento “A agência meteorológica do Japão” atua como agente do verbo “chegar (chegou)”, mas não foi anotado pelo sistema. Por último, na categoria *Política*, o segmento “pela Mesa Diretora do Senado” representa o agente da passiva, portanto, é difícil para o classificador identifica-lo como agente.

Categoria	Sentença
Cotidiano	Segundo o presidente, a prioridade é a realização de obras nas regiões metropolitanas de grandes centros urbanos.
Esportes	Com o resultado, o Brasil **está** na liderança do grupo B, perto da classificação para a próxima fase do campeonato.
Mundo	A agência meteorológica do Japão **chegou** a emitir alerta de Tsunami, mas o cancelou uma hora após.
Política	Amanhã **será decidido**, pela Mesa Diretora do Senado, se a quarta representação será encaminhada ao Conselho de Ética.

Tabela 9: Falsos negativos do aspecto WHO_AGENT

Falsos positivos

Na Tab. 10, mostram-se algumas das sentenças cujos aspectos foram identificados automaticamente pelo sistema APS + Regras, mas que não foram anotadas manualmente. Observa-se que nas categorias *Cotidiano*, *Esportes* e *Política*, os segmentos “Lula”, “Maradona” e “Ele”, respectivamente, foram identificados corretamente como WHO_AGENT, porém não foram anotadas manualmente. Isso pode ter sido erro de anotação humana. Na sentença da categoria *Mundo*, o segmento “O furação Dean” não representa uma organização/pessoa; o sistema errou ao considerar erroneamente “Dean” uma pessoa.

Categoria	Sentença
Cotidiano	Depois de as vaias, <aspect APS="WHO_AGENT">Lula</aspect> **desistiu** de declarar abertos os jogos, como estava planejado.
Esportes	<aspect APS="WHO_AGENT">Maradona</aspect> **voltou** a ter problemas de saúde em o fim de semana e foi internado novamente em um hospital em Buenos Aires.
Mundo	<aspect APS="WHO_AGENT">O furacão Dean</aspect> **passou** por a costa sul de a Jamaica, inundando a capital e espalhando árvores e telhados.
Política	<aspect APS="WHO_AGENT">Ele</aspect> **disse que** é ideal que se crie uma comissão de três relatores para os processos em conjunto.

Tabela 10*:* Falsos positivos do aspecto WHO_AGENT

4.2. WHO_AFFECTED

Segundo Rassi et al. (2013), o aspecto WHO_AFFECTED é definido como *“a entidade (pessoa ou organização) que sofre efeitos de um fato/*evento”. Ao contrário do aspecto WHO_AGENT, o aspecto WHO_AFFECTED é representado pelo sujeito gramatical paciente da oração. Na Fig. 13, apresenta-se uma sentença anotada com o aspecto WHO_AFFECTED. Nota-se que o sujeito

paciente é definido pelo segmento “17 pessoas”, que está relacionado semanticamente ao verbo “morrer (morreram)” e representa a entidade pessoa.

17 pessoas morreram após a queda de um avião na República Democrática do Congo.

Figura 13: *Sentença anotada com microaspecto WHO_AFFECTED*

Para identificar o aspecto WHO_AFFECTED, foram utilizados os sistemas APS e APS + Regras. No entanto, não foi utilizado o sistema REMBRANDT, por este ser incapaz de identificar o sujeito paciente da oração. O sistema foi testado sobre o córpus CSTNews com um total de 60 sentenças anotadas manualmente com o aspecto WHO_AFFECTED (ver Tab. 11). Observa-se que o aspecto não está bem distribuído entre todas as categorias, sendo *Esportes* a categoria com menos dados anotados.

Categoria	Frequência
Cotidiano	10
Esportes	6
Mundo	26
Política	18
Total	**60**

Tabela 11: Distribuição do *microaspecto* WHO_AFFECTED por categoria

Na Tab. 12, apresentam-se os resultados dos sistemas avaliados. Observa-se que o sistema APS + Regras superou o sistema APS. O uso de regras para identificar a existência de uma entidade pessoa/organização dentro de um segmento teve êxito. Porém, a cobertura diminuiu de 0.767 a 0.417. Isso quer dizer que, assim como para o aspecto WHO_AGENT, o sistema APS estava acertando por acaso. Na Tab. 13, mostra-se a matriz de confusão do sistema APS + Regras. Nota-se que, das 60 sentenças anotadas, só 25 sentenças foram anotadas pelo sistema APS + Regras corretamente. A seguir, analisam-se algumas sentenças identificadas como “falsos negativos” e “falsos positivos” do sistema APS + Regras.

WHO_AFFECTED	Cobertura	Precisão	F1	Acurácia
APS	0.767	0.203	0.321	0.394
APS + Regras	**0.417**	**0.368**	**0.391**	**0.758**

Tabela 12: Resultados para o *microaspecto* WHO_AFFECTED

WHO_AFFECTED	Verdadeiro (P)	Falso (P)	**Classe**	Cobertura	Precisão	F1	Acurácia
Verdadeiro	25	35	Sim	0.417	0.368	0.391	0.758
Falso	43	219	Não	0.836	0.862	0.849	

Tabela 13: Matriz de confusão do aspecto WHO_AFFECTED

Falsos negativos

Na Tab. 14, mostram-se algumas das sentenças cujos aspectos não foram identificados automaticamente pelo sistema APS + Regras, mas que foram anotadas manualmente. Observa-se claramente que, na sentença da categoria *Cotidiano*, o classificador não conseguiu identificar o segmento “200 pessoas” como sujeito paciente da oração, relacionado ao verbo

"vitimar (vitimado)". Isso é um erro do sistema APS. Caso similar ocorre na sentença da categoria *Mundo* e *Política*. Já na sentença da categoria *Esportes*, acontece um problema de sujeito oculto. Assim, a entidade "Maradona", relacionada ao verbo "internar (internado)", não foi identificada.

Categoria	Sentença
Cotidiano	Em o acidente, o avião passou por a pista de Congonhas com velocidade acima de o normal, atravessou uma avenida e atingiu um prédio, **vitimando** 200 pessoas.
Esportes	Maradona voltou a ter problemas de saúde em o fim de semana e **foi internado** novamente em um hospital em Buenos Aires.
Mundo	O terremoto **deixou** 9 pessoas mortas (todos idosos) e mais de 700 feridos, além de casas e viadutos destruídos.
Política	Cristovam Buarque e Luciano Bivar **têm**, cada um, 1% dos votos.

Tabela 14: Falsos negativos do aspecto WHO_AGENT

Falsos positivos

Na Tab. 15, apresentam-se algumas das sentenças que foram identificadas automaticamente pelo sistema APS + Regras, mas que não foram anotadas manualmente. Nota-se que na sentença da categoria *Cotidiano*, a organização "A secretaria da Fazenda" foi identificado corretamente, porém, essa sentença não foi anotada manualmente. Caso similar ocorre com a sentença da categoria *Esportes*. Um problema do classificador APS + Regras é que, na maioria dos casos, anotam-se segmentos posicionados à esquerda do verbo alvo. Isso pode ser observado nas sentenças das categorias *Mundo* e *Política*, em que os segmentos "de a polícia" e "por a bancada de o PT", respectivamente, foram erroneamente classificados.

Categoria	Sentença
Cotidiano	<aspect APS="WHO_AFFECTED">A Secretaria da Fazenda</aspect> também **foi atingida** por uma bomba.
Esportes	Ronaldinho fez uma sequência de dribles em o Equador e **cruzou** <aspect APS="WHO_AFFECTED">para Elano</aspect>, que fez o quarto gol.
Mundo	Depois de isso, fez quatro cirurgias plásticas para **escapar** <aspect APs="WHO_AFFECTED">de a polícia</aspect>.
Política	A unificação **foi proposta** <aspect APS="WHO_AFFECTED">por a bancada de o PT</aspect> e tem apoio de o PSOL.

Tabela 15: Falsos positivos do *microaspecto* WHO_AFFECTED

4.3. WHEN

Segundo Rassi et al. (2013), o aspecto WHEN é definido como "*a data/período de tempo (estritamente temporal) de ocorrência de um fato/evento*". Na Fig. 14, ilustra-se uma sentença anotada com o aspecto WHEN. Observa-se que o segmento "na quarta-feira, 16" representa uma data.

Um homem suspeito de ter roubado o relógio Rolex do apresentador de televisão Luciano Huck foi detido na quarta-feira, 16, em Taboão da Serra, na Grande São Paulo.

Figura 14: Sentença anotada com *microaspecto* WHEN

Foram testados 4 tipos de sistemas: APS, REMBRANDT, APS + REMBRANDT, APS + Regras. Os sistemas foram testados sobre o córpus CSTNews com um total de 75 sentenças anotadas manualmente com o aspecto WHEN (ver Tab. 16). Observa-se que o aspecto está bem distribuído entre todas as categorias. Revisando as tabelas do Apêndice A, nota-se que o aspecto WHEN foi definido para todas as categorias.

Categoria	Frequência
Cotidiano	25
Esportes	14
Mundo	22
Política	14
Total	**75**

Tabela 16: Distribuição do *microaspecto* WHEN por categoria

Os resultados dos 4 sistemas são apresentados na Tab. 17. Nota-se que os melhores resultados foram obtidos pelo sistema APS + Regras. Na Tab. 18, apresenta-se a matriz de confusão do melhor sistema (APS + Regras) para o *microaspecto* WHEN. Mesmo assim, a precisão é relativamente baixa (0.504), por causa da grande quantidade de "falsos positivos". A seguir, analisam-se algumas sentenças identificadas como "falsos negativos" e "falsos positivos".

WHEN	Cobertura	Precisão	F1	Acurácia
APS	0.693	0.477	0.565	0.752
REMBRANDT	0.547	0.719	0.621	0.845
APS + REMBRANDT	0.840	0.492	0.621	0.761
APS + Regras	**0.947**	**0.504**	**0.657**	**0.770**

Tabela 17: Resultados para o *microaspecto* WHEN

WHEN	Verdadeiro (P)	Falso (P)	**Classe**	Cobertura	Precisão	F1	Acurácia
Verdadeiro	71	4	Sim	0.947	0.504	0.657	0.770
Falso	70	177	Não	0.717	0.978	0.827	

Tabela 18: Matriz de confusão do *microaspecto* WHEN

Falsos negativos

Na Tab. 19, mostram-se algumas das sentenças cujos aspectos não foram identificados pelo sistema APS + Regras. Observa-se que na sentença da categoria *Cotidiano*, os segmentos "às 8h", "às 9h" e "meia hora depois" não foram identificados. Isso quer dizer que deve ser criada uma regra para identificar tempo em formato de horas. Na sentença da categoria *Política*, não foi identificado o advérbio de tempo "Amanhã". Tentou-se solucionar esse problema criando uma regra que identifica advérbios de tempo isolados, ou seja, que não foram antecedidos pela preposição (como referência, ver regra nº 4 da Fig. 21, no Apêndice B). Porém, a regra identificava muitas sentenças que não foram anotadas manualmente, gerando muitos "falsos positivos". Por causa disso, optou-se por não a utilizar.

Categoria	Sentença
Cotidiano	A Companhia de Engenharia de Tráfego (CET) anunciou que o índice de congestionamento era de 54 quilômetros às 8h, 113 km às 9h e 110 km meia hora depois, valores bem acima de as médias para os horários, que eram de 36, 82 e 76 quilômetros respectivamente.
Política	Amanhã será decidido, por a Mesa Diretora de o Senado, se a quarta representação será encaminhada a o Conselho de Ética.

Tabela 19: Falsos negativos do *microaspecto* WHEN

Falsos positivos

Na Tab. 20, mostram-se algumas das sentenças que o sistema APS + Regras não devia ter identificado. As sentenças da categoria *Cotidiano, Mundo* e *Política* foram anotadas corretamente, mas não foram anotadas manualmente. Já na sentença da categoria *Esportes*, o classificador errou totalmente ao classificar o segmento "a 0" como tempo.

Categoria	Sentença
Cotidiano	<aspect APS="WHEN">Antes de a festa</aspect>, ele visitou a Vila Olímpica e conversou com atletas de vários países de o mundo que estavam lá e ouviu de vários de eles elogios sobre a qualidade de o que o Brasil estava oferecendo em a Vila Olímpica.
Esportes	A equipe brasileira, comandada por Bernardinho, **venceu** a Finlândia por 3 sets <aspect APS="WHEN">a 0</aspect>, em Tampere (FIN), mantendo sua invencibilidade em a Liga Mundial de Vôlei-06.
Mundo	Esta proposta será debatida por a ONU <aspect APS="WHEN">hoje</aspect> ou amanhã.
Política	O grupo criminoso desviou <aspect APS="WHEN">desde 2004</aspect> cerca de R$ 70 milhões de os cofres públicos.

Tabela 20: Falsos positivos do aspecto WHEN

4.4. WHERE

Segundo Rassi et al. (2013), o aspecto WHERE é definido como "*a localização geográfica ou física de um fato/evento*". Na Fig. 15, ilustra-se uma sentença anotada com o aspecto WHERE. Observa-se que o segmento "em São Paulo" representa um local geográfico.

> Na sexta-feira, em encontro com sindicalistas em São Paulo, Lula disse que venceria a eleição no primeiro turno - tendência apontada pelas pesquisas.

Figura 15: Sentença anotada com *microaspecto* WHERE

Da mesma forma que para o *microaspecto* WHEN, foram testados 4 tipos de sistemas: APS, REMBRANDT, APS + REMBRANDT, APS + Regras. Os sistemas foram testados sobre o córpus CSTNews com um total de 56 sentenças anotadas manualmente com o aspecto WHERE (ver Tab. 21). Nota-se que as sentenças da categoria *Política* não foram anotadas. Portanto, existe a possibilidade de se gerar uma grande quantidade de "falsos positivos".

Categoria	Frequência
Cotidiano	22
Esportes	9
Mundo	25
Política	0
Total	**56**

Tabela 21: Distribuição do *microaspecto* WHERE por categoria

Os resultados dos 4 sistemas são apresentados na Tab. 22. Nota-se que os melhores resultados foram obtidos pelo sistema APS + Regras. Na Tab. 23, apresenta-se a matriz de confusão do melhor sistema (APS + Regras) para o *microaspecto* WHERE. Mesmo assim, a precisão é baixa (0.474) por causa da grande quantidade de "falsos positivos". A seguir, analisam-se algumas sentenças identificadas como "falsos negativos" e "falsos positivos".

WHERE	Cobertura	Precisão	F1	Acurácia
APS	0.679	0.447	0.539	0.798
REMBRANDT	0.804	0.425	0.556	0.776
APS + REMBRANDT	0.946	0.363	0.525	0.702
APS + Regras	**0.804**	**0.474**	**0.596**	**0.811**

Tabela 22: Resultados para o *microaspecto* WHERE

WHERE	Verdadeiro (P)	Falso (P)	**Classe**	Cobertura	Precisão	F1	Acurácia
Verdadeiro	45	11	Sim	0.804	0.474	0.596	0.811
Falso	50	216	Não	0.812	0.952	0.876	

Tabela 23: Matriz de confusão do *microaspecto* WHERE

Falsos negativos

Na Tab. 24, mostram-se algumas das sentenças cujos aspectos não foram identificados pelo sistema APS + Regras. Nota-se que o segmento "para a Bahia", na categoria *Cotidiano*, contém a preposição "para" em lugar da preposição "em" (ver Fig. 22, no Apêndice B). Da mesma forma, o segmento "Norris Hall", na categoria *Mundo*, não está associado à preposição "em".

Categoria	Sentença
Cotidiano	A família pediu que o corpo fosse levado diretamente para a Bahia.
Mundo	Pouco depois, Norris Hall, edifício de a engenharia, foi alvo de outro ataque a tiros.

Tabela 24: Falsos negativos do *microaspecto* WHERE

Falsos positivos

Na Tab. 25, mostram-se algumas das sentenças cujos aspectos o sistema APS + Regras não devia ter identificado como WHERE. Tanto na categoria *Cotidiano* quanto na categoria *Política*, o sistema classifica corretamente os segmentos "em o Palácio de Aclamação" e "em Rondônia" como expressões de lugar, contudo, não foram anotadas manualmente. Já na categoria *Esportes* e *Mundo*, o sistema APS identificou erroneamente os segmentos "em o salto com vara" e "em um de seus transformadores".

Categoria	Sentença
Cotidiano	O velório será <aspect APS="WHERE"> em o Palácio da Aclamação. </aspect>
Esportes	A brasileira Fabiana Murer **conquistou** a medalha de ouro <aspect APS="WHERE">em o salto com vara</aspect> a o saltar 4m60, um novo recorde pan-americano, 20 cm a mais que sua antiga marca.
Mundo	A maior usina nuclear de o mundo **teve** incêndio <aspect APS="WHERE">em um de seus transformadores</aspect>, mas o fogo foi controlado e não houve vazamento de radiação.
Política	Algumas pessoas pertencem a o alto escalão político <aspect APS="WHERE">em Rondônia.</aspect>.

Tabela 25: Falsos positivos do aspecto WHERE

4.5. WHY

Segundo Rassi et al. (2013), o aspecto WHY é definido como "*uma explicação do porquê um fato/evento acontece (ou aconteceu)*". Na Fig. 16, mostra-se uma sentença anotada com o aspecto WHY. Observa-se que o segmento "porque" representa uma expressão explícita de causa ou motivo.

O crescimento nas autuações de contribuintes que caíram na malha fina se deu porque os auditores passaram a contar com programas mais modernos de computadores que analisam todas as irregularidades fiscais dos contribuintes, inclusive de anos anteriores, e não mais por grupos de infrações.

Figura 16: Sentença anotada com *microaspecto* WHY

No total, foram testados 2 tipos de sistemas: APS e APS + Regras. O sistema REMBRANDT foi descartado por não possuir uma categoria equivalente ao *microaspecto* WHY (ver Tab. 3). Os sistemas foram testados sobre o córpus CSTNews com um total de 32 sentenças anotadas manualmente com o aspecto WHY (ver Tab. 26). Nota-se que a quantidade de aspectos anotados foi pouca, sendo *Cotidiano* e *Esportes* as categorias com menos dados anotados.

Categoria	Frequência
Cotidiano	4
Esportes	3
Mundo	14
Política	11
Total	**32**

Tabela 26: Distribuição do *microaspecto* WHY por categoria

Na Tab. 27, mostram-se os resultados dos sistemas avaliados. Nota-se que o sistema APS + Regras ganhou do sistema APS por uma considerável diferença em todas as métricas. Mesmo assim, a cobertura é baixa (0.466). Na Tab. 28, apresenta-se a matriz de confusão do melhor sistema (APS + Regras) para o *microaspecto* WHY. Pode-se observar que, mesmo com o auxílio das regras, encontram-se 17 "falsos negativos" de 32 sentenças anotadas. A seguir, analisam-se algumas sentenças identificadas como "falsos negativos" e "falsos positivos".

WHY	Cobertura	Precisão	F1	Acurácia
APS	0.156	0.500	0.238	0.901
APS + Regras	**0.469**	**0.789**	**0.588**	**0.935**

Tabela 27: Resultados para o *microaspecto* WHY

WHY	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	15	17	Sim	0.469	0.789	0.588	0.935
Falso	4	286	Não	0.986	0.944	0.966	

Tabela 28: Matriz de confusão do *microaspecto* WHY

Falsos negativos

Na Tab. 29, mostram-se algumas das sentenças cujos aspectos não foram identificados pelo sistema APS + Regras, mas que foram anotadas manualmente. Nota-se que, nas sentenças da categoria *Esportes* e *Mundo*, não existe uma expressão de causa explícita, dessa maneira, as regras não conseguiram classificar as sentenças como WHY (ver Fig. 23, no Apêndice B). Por outro lado, a sentença da categoria *Política* não representa uma causa explícita ou implícita, portanto, não devia ser anotada manualmente.

Categoria	Sentença
Esportes	Maradona voltou a ter problemas de saúde em o fim de semana e foi internado novamente em um hospital em Buenos Aires.
Mundo	Duas mulheres de 80 anos morreram no desmoronamento de suas casas.
Política	Tem candidato aí com o salto 15, eu vou nas sandálias da humildade, disse Alckmin, apontando a arrogância do adversário.

Tabela 29: Falsos negativos do aspecto WHY

Falsos positivos

Na Tab. 30, mostram-se algumas das sentenças que o sistema APS + Regras não devia ter identificado como WHY. Nota-se que, na sentença da categoria *Cotidiano*, identificou-se a expressão "pois", porém, a expressão não denota uma causa, estritamente falando. Já na categoria *Política*, a expressão "porque" representa uma causa explícita, portanto, devia ser anotada manualmente.

Categoria	Sentença
Cotidiano	Seu discurso foi focado em as questões ambientais, citando que a produção de biocombustíveis não afeta a segurança alimentar, <aspect APS="WHY">pois</aspect> a cana-de-açúcar ocupa apenas 1 % de as terras agricultáveis de o Brasil.
Política	Eu não moverei uma palha contra a oposição <aspect APS="WHY"> porque</aspect> vocês moverão um paiol inteiro, afirmou o presidente Luiz Inácio Lula da Silva, candidato à reeleição por o PT, sobre os ataques de seus adversários.

Tabela 30: Falsos positivos do aspecto WHY

4.6. HOW

Segundo Rassi et al. (2013), o aspecto HOW é definido como "*o modo como um fato/evento ocorre*". Na Fig. 17, mostra-se uma sentença anotada com o aspecto HOW. Observa-se que o segmento "em três tentativas" indica a maneira em que a esportista "Fabiana" conseguiu a "medalha de ouro".

Fabiana conseguiu o ouro em três tentativas.

Figura 17: Sentença anotada com *microaspecto* HOW

O único sistema testado foi o APS. O sistema REMBRANDT foi desconsiderado por não ter uma categoria equivalente ao aspecto HOW. Como já foi dito, não foi possível criar regras porque existem muito poucas sentenças anotadas e não foi possível identificar padrões. O sistema foi testado sobre o córpus CSTNews com um total de 9 sentenças anotadas manualmente com o aspecto HOW (ver Tab. 31). Nota-se que as sentenças da categoria *Esportes* e *Mundo* não foram anotadas. Ao revisar as tabelas do Apêndice A, percebe-se que HOW não foi definido como *microaspecto* nas categorias *Esportes* e *Mundo*.

Categoria	Frequência
Cotidiano	6
Esportes	0
Mundo	0
Política	3
Total	**9**

Tabela 31: Distribuição do *microaspecto* HOW por categoria

Na Tab. 32, mostra-se a matriz de confusão do sistema APS para o *microaspecto* HOW. Observa-se que a medida F1 foi baixa (0.040) em função da grande quantidade de "falsos negativos e positivos". A seguir, analisam-se algumas sentenças identificadas como "falsos negativos" e "falsos positivos".

HOW	Verdadeiro (P)	Falso (P)	**Classe**	Cobertura	Precisão	F1	Acurácia
Verdadeiro	1	8	Sim	0.111	0.024	0.040	0.851
Falso	40	273	Não	0.872	0.972	0.919	

Tabela 32: Matriz de confusão do *microaspecto* HOW

Falsos negativos

Na Tab. 33, mostram-se algumas das sentenças que o sistema APS não devia ter identificado como HOW. Nota-se que, nas sentenças da categoria *Cotidiano* e *Política*, os segmentos "com bombas e tiros" e "em dois turnos" não foram identificados pelo sistema.

Categoria	Sentença
Cotidiano	Os bandidos atacaram agências bancárias, bases policiais e prédios públicos com bombas e tiros.
Política	Para ser aprovada, a PEC precisa ser votada em dois turnos da Câmara.

Tabela 33: Falsos negativos do aspecto HOW

Falsos positivos

Na Tab. 34, apresentam-se algumas das sentenças cujos aspectos o sistema APS não identificou automaticamente como HOW. Nota-se que os segmentos marcados nas sentenças das categorias *Cotidiano* e *Política* foram erradamente identificados como HOW, por seguir o padrão "com" e "como" no começo dos segmentos.

Categoria	Sentença
Cotidiano	A falha em o reversor -- mecanismo que ajuda o avião a frear -- foi detectada por o sistema de a aeronave, que continuou **voando** em os dias seguintes <aspect APS="HOW">com o reversor desligado</aspect>.
Política	O assunto surgiu depois que seu advogado questionou a legitimidade de o colegiado e de a própria Polícia Federal em **investigar** o caso, uma vez que, <aspect APS="HOW">como senador, Renan gozaria de foro especial</aspect>.

Tabela 34: Falsos positivos do aspecto HOW

4.7. SITUATION

Segundo Rassi et al. (2013), o aspecto SITUATION é definido como "*uma ocasião em que ocorreu um fato/evento. Envolve uma transação, um campeonato, um compromisso ou outros tipos de situação em uma data ou local inespecíficos.*" Na Fig. 18, mostra-se uma sentença anotada com o aspecto SITUATION. Observa-se claramente que o segmento "Jogos Pan-Americanos" indica campeonato ou competição.

A equipe de revezamento 4x200 metros livre conquistou nesta terça-feira a segunda medalha de ouro da natação brasileira nos Jogos Pan-Americanos.

Figura 18: Sentença anotada com *microaspecto* SITUATION

O sistema escolhido foi o REMBRANDT. O sistema APS foi descartado por não ter um papel semântico equivalente ao aspecto SITUATION. O sistema foi testado sobre o córpus CSTNews com um total de 13 sentenças anotadas manualmente (ver Tab. 35). Nota-se que conjunto de sentenças anotadas na categoria *Mundo* é bem escasso. Cabe ressaltar que o aspecto SITUATION, nas categorias *Cotidiano* e *Política*, ocorre como *macroaspecto*, portanto, não tem sentenças anotadas com SITUATION como *microaspecto* (ver Apêndice A).

Categoria	Frequência
Cotidiano	0
Esportes	11
Mundo	2
Política	0
Total	**13**

Tabela 35: Distribuição do *microaspecto* SITUATION por categoria

Na Tab. 36, mostra-se a matriz de confusão do sistema REMBRANDT para o *microaspecto* SITUATION. Em primeiro lugar, observa-se que, no total, foram testadas 154 sentenças das categorias *Esportes* e *Mundo*. Em segundo lugar, nota-se que a cobertura foi baixa (0.231), enquanto a precisão foi alta (0.750). A seguir, analisam-se algumas sentenças identificadas como "falsos negativos" e "falsos positivos".

SITUATION	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	3	10	Sim	0.231	0.750	0.353	0.929
Falso	1	140	Não	0.993	0.933	0.962	

Tabela 36: Matriz de confusão do *microaspecto* SITUATION

Falsos negativos

Na Tab. 37, mostram-se algumas das sentenças que o sistema REMBRANDT não conseguiu identificar como SITUATION. Nota-se que o segmento "Liga Mundial de Vôlei-06", na categoria *Esportes*, é uma entidade mencionada que indica um evento competitivo, portanto, devia ser identificada pelo sistema. Por outro lado, o segmento "em esta batalha", na categoria *Mundo*, não é uma entidade mencionada (ou nome próprio), portanto, o sistema não conseguiu identifica-la.

Categoria	Sentença
Esportes	A equipe brasileira, comandada por Bernardinho, venceu a Finlândia por 3 sets a 0, em Tampere (FIN), mantendo sua invencibilidade na Liga Mundial de Vôlei-06.
Mundo	Nesta batalha, 15 soldados israelenses morreram ao serem atingidos por um míssil.

Tabela 37: Falsos negativos do aspecto SITUATION

Falsos positivos

Na Tab. 38, apresenta-se a única sentença que o sistema REMBRANDT identificou automaticamente como SITUATION, mas que não foi anotada manualmente. Nota-se claramente que o segmento "Jogos Olímpicos de Pequim" indica uma competição, mas não foi anotado manualmente (erro de anotação).

Categoria	Sentença
Esportes	A ginasta Jade Barbosa foi escolhida, em votação na Internet, para ser a representante do Brasil no revezamento da tocha dos <aspect EM= "ACONTECIMENTO|ORGANIZADO">Jogos Olímpicos de Pequim </aspect>

Tabela 38: Falsos positivos do aspecto SITUATION

4.8. SCORE

Segundo Rassi et al. (2013), o aspecto SCORE é definido como "o *resultado numérico de um fato/evento (score, tempo, distância, etc., sobretudo relativo a esportes)*". Na Fig. 19, ilustra-se uma sentença anotada com o aspecto SCORE. Observa-se que o segmento "3 a 0" indica o resultado em gols de um jogo de futebol.

A seleção brasileira, sob direção de Dunga, conquistou o oitavo título da Copa América, goleando a Argentina por 3 a 0.

Figura 19: Sentença anotada com *microaspecto* SCORE

O único sistema testado foi o APS + Regras. É importante lembrar que o aspecto SCORE não tem equivalência com algum papel semântico, portanto, só foram criadas regras manuais e integradas ao sistema. Já o sistema REMBRANDT foi desconsiderado por não ter uma categoria

equivalente ao aspecto SCORE. O sistema foi testado sobre o córpus CSTNews com um total de 10 sentenças anotadas manualmente (ver Tab. 39). Nota-se que o *microaspecto* SCORE só ocorre nas sentenças da categoria *Esportes*. De maneira igual, no Apêndice A, observa-se que o aspecto SCORE foi definido só para *Esportes*.

Categoria	Frequência
Cotidiano	0
Esportes	10
Mundo	0
Política	0
Total	**10**

Tabela 39: Distribuição do *microaspecto* SCORE por categoria

Na Tab. 40, mostra-se a matriz de confusão do sistema APS + Regras para o aspecto SCORE. Observa-se que as regras manuais tiveram um ótimo desempenho, identificando todas as sentenças anotadas manualmente. Cabe ressaltar que as regras foram cridas sobre as 10 sentenças anotadas, por isso o resultado é excelente.

SCORE	Verdadeiro (P)	Falso (P)	**Classe**	Cobertura	Precisão	F1	Acurácia
Verdadeiro	10	0	Sim	1.000	1.000	1.000	1.000
Falso	0	312	Não	1.000	1.000	1.000	

Tabela 40: Matriz de confusão do *microaspecto* SCORE

4.9. Resultados do sistema AM

Diferentemente dos sistemas testados anteriormente, este classificador foi *treinado* e *testado* com as 322 sentenças (ou instâncias) do córpus CSTNews anotadas com aspectos. A estratégia de divisão dos conjuntos de treinamento e teste foi de *estratificação*, já que se garante que haja as mesmas proporções de classes dentro de cada subconjunto. A ideia de se usar córpus estratificado é de amenizar o problema de "desbalanceamento de classes", que pode influenciar no desempenho do classificador (Blake and Merz, 1998). Portanto, o córpus foi 10 vezes (iterações) estratificado, sendo que, para cada iteração, a divisão do córpus foi de 70% para o conjunto de treinamento (225 instâncias) e 30% para o conjunto de teste (97 instâncias).

A identificação de *microaspectos* é um problema de classificação multirrótulo. Neste trabalho aplica-se o método de transformação de problemas (Tsoumakas e Katakis, 2007), que visa transformar o problema de classificação multirrótulo em um conjunto de problemas de classificação binária. Portanto, foram gerados vários classificadores binários, sendo escolhidos somente 8 classificadores, um para cada *microaspecto* (WHO_AGENT, WHO_AFFECTED, WHERE, WHEN, WHY, HOW, SITUATION e SCORE). Cabe ressaltar que o aspecto GOAL não foi considerado.

No total, foram definidos 6 tipos de atributos semânticos e morfossintáticos (ver Tab. 41), fornecidos pelo PALAVRAS no formato *flat* (ver Fig. 4, na Seção 2.2). Cada atributo é representado por unigramas "(1, 1)", bigramas "(2, 2)" e bigramas + trigramas "(2, 3)". Assim, para cada um dos 8 *microaspectos*, cria-se um classificador resultado da representação (unigramas, bigrama, bigrama + trigrama) de cada um dos 6 tipos de atributos (ver Apêndice

C). Por exemplo, o classificador denominado "(1, 1) POS", foi criado com base em todos os unigramas "(1, 1)" das classes gramaticais (POS) de todas as palavras dos textos-fonte. No total, foram criados 144 classificadores.

Tipo de atributo	Notação
bag of words	bag_of_words
lematização	lemmas
POS (*part-of-speech*)	POS
tags-semânticos	semantic
lematização + POS	lemmas+POS
POS + tags-semânticos	POS+semantic

Tabela 41: Atributos definidos

Como já foi dito, a técnica de AM supervisionada usada foi o SVM. No trabalho futuro, serão utilizadas outras técnicas como Árvores de Decisão (Breiman et al., 1984), Redes neurais (Haykin, 1998) ou Redes bayesianas (Mitchell, 1997). A avaliação de cada classificador foi feita conforme as métricas estatísticas obtidas da matriz de confusão: "Precisão", "Cobertura", "F1" e "Acurácia". O resultado final é a **média** dos valores obtidos em cada uma das 10 iterações do córpus estratificado. Na Tab. 42, mostram-se os resultados dos melhores classificadores para cada *microaspecto*. No Apêndice C, mostram-se os resultados de todos os classificadores.

Microaspecto	Sentença	Cobertura	Precisão	F1	Acurácia
WHO_AGENT	(2, 3) POS+semantic	0.538	0.636	**0.583**	0.691
WHO_AFFECTED	(1, 1) lemmas	0.222	1.000	**0.364**	0.854
WHEN	(1, 1) semantic	0.522	0.750	**0.615**	0.845
WHERE	(2, 3) POS+semantic	0.471	0.615	**0.533**	0.856
WHY	(2, 3) POS+semantic	0.200	0.500	**0.286**	0.897
HOW	(1, 1) bag_of_words	0.250	1.000	**0.400**	0.938
SITUATION	(1, 1) lemmas+POS	0.333	1.000	**0.500**	0.959
SCORE	Todos	0.000	0.000	**0.000**	0.000

Tabela 42: Melhores resultados dos *microaspectos* usando AM

Observa-se que o desempenho dos classificadores, em termos de F1, é baixo na maioria dos casos, a exceção dos aspectos WHO_AGENT, WHEN e WHERE. Isso se deve às poucas instâncias anotadas. O melhor resultado foi obtido pelo classificador "*(1, 1) semantic*" para o aspecto WHEN. O pior resultado foi obtido por todas as combinações de classificadores para o aspecto SCORE. A causa do aspecto SCORE ter ido mal é a pouca quantidade de sentenças anotadas (10) no córpus. Nota-se, também, que a maioria dos melhores classificadores é representada por unigramas "(1, 1)". Por último, o atributo mais representativo é "*(2, 3) POS+semantic*", alcançando os melhores resultados nos aspectos WHO_AGENT, WHERE e WHY.

A seguir, apresenta-se um quadro com os melhores resultados dos 4 sistemas avaliados para cada *microaspecto* (ver Tab. 43). É importante ter em consideração que os sistemas APS, APS + Regras e REMBRANDT foram testados sobre 322 sentenças anotadas do córpus CSTNews. Já o sistema AM só foi testado com 97 das 322 sentenças.

Microaspecto	Classificador	Cobertura	Precisão	F1	Acurácia
WHO_AGENT	APS + Regras	0.592	0.664	**0.626**	0.624
WHO_AFFECTED	APS + Regras	0.417	0.368	**0.391**	0.758
WHEN	APS + Regras	0.947	0.504	**0.657**	0.770
WHERE	APS + Regras	0.804	0.474	**0.596**	0.811
WHY	APS + Regras	0.469	0.789	**0.588**	0.935
HOW	AM (1, 1) bag_of_words	0.250	1.000	**0.400**	0.938
SITUATION	AM (1, 1) lemmas+POS	0.333	1.000	**0.500**	0.959
SCORE	APS + Regras	1.000	1.000	**1.000**	1.000

Tabela 43: Melhores resultados dos sistemas propostos

Nota-se que os aspectos WHO_AGENT, WHEN, WHERE, WHY e SCORE alcançaram os melhores resultados com o classificador/sistema APS + Regras. Já o aspecto WHO_AFFECTED obteve o pior resultado. Observa-se claramente que o sistema APS + Regras abrange a maioria dos *microaspectos*.

5. Conclusões

Nesse relatório, foram apresentados o processo e os resultados da "identificação automática de *microaspectos*". No total, foram avaliados dois tipos de abordagens: usando sistemas propostos (APS, APS + Regras, REMBRANDT) e usando técnicas de AM. As abordagens foram testadas sobre o mesmo córpus (CSTNews). O sistema APS está baseado no anotador de papéis semânticos de Alva-Manchego (2003). O sistema APS + Regras se baseia no conjunto de regras manuais integradas ao sistema APS. Já o sistema REMBRANDT está baseado no reconhecedor de entidades mencionadas de Cardoso (2008). Por último, foram criados vários classificadores usando técnicas de AM, com base na combinação de atributos morfossintáticos (*bag-of-words*, lematização e POS) e semânticos (tags-semânticos).

A abordagem usando sistemas foi testada sobre um total de 322 sentenças anotadas manualmente dos sumários multidocumento do CSTNews. Os resultados mostraram que o sistema APS + Regras foi o melhor para a maioria dos *microaspectos* (WHO_AGENT, WHEN, WHO_AFFECTED, WHERE, WHY e SCORE). Isso quer dizer claramente que as regras melhoraram o desempenho do sistema APS. Já o sistema APS só conseguiu o melhor resultado para o *microaspecto* HOW. Cabe ressaltar que, devido a pouca quantidade de dados anotados, não foram criadas regras para o *microaspecto* HOW. Da maneira igual, o sistema REMBRANDT só obteve um resultado bom para o *microaspecto* SITUATION. Por último, é importante ressaltar os problemas identificados no processo de identificação:

1. Em algumas ocasiões, o sistema APS teve problemas ao não conseguir classificar alguns papéis semânticos ou ao classificar papéis de maneira errada, afetando o desempenho do sistema APS + Regras.
2. Algumas sentenças não foram analisadas sintaticamente pelo *parser* PALAVRAS, aumentando a quantidade de "falsos negativos".
3. O sistema REMBRANDT só identificava entidades nomeadas escritas em caixa alta (a exceção das expressões temporais), causando um baixo desempenho do sistema.

Diferentemente da abordagem usando sistemas, a abordagem utilizando técnicas de AM foi testada com apenas 30% do córpus CSTNews, ou seja, um total de 97 sobre 322 instâncias. Assim, pode-se dizer que o baixo desempenho dos classificadores usando AM se deve à pouca quantidade de instâncias de treino e teste. Acredita-se que a existência de mais instâncias/sentenças no córpus possa melhorar os resultados dos classificadores.

Um dos grandes fatores pelo qual os resultados não foram relativamente altos é a anotação de aspectos do córpus CSTNews (Rassi et al., 2008). Olhando-se para os "falsos positivos" apresentados na seção dos resultados, percebe-se, em várias ocasiões, que os sistemas anotaram automaticamente sentenças que não foram anotadas manualmente. Na Tab. 21, observa-se claramente que o aspecto WHERE não ocorre na categoria *Política*. Portanto, é muito provável que o sistema identifique sentenças que expressem lugar, que não foram anotadas manualmente. Por outro lado, é factível fazer uma análise dos "falsos negativos" para, assim, aprimorar e melhorar o desempenho dos sistemas correspondentes.

Trabalhos futuros incluem a identificação de *macroaspectos*, que, juntamente com os *microaspectos*, devem subsidiar a exploração de métodos de sumarização com base em aspectos.

Agradecimentos

Os resultados apresentados neste relatório foram obtidos no âmbito do convênio universidade-empresa intitulado "Processamento Semântico de Textos em Português Brasileiro", financiado pela Samsung Eletrônica da Amazônia Ltda., nos termos da legislação federal brasileira nº 8.248/91

Referências

Alva-Manchego, Fernando. 2013. "Anotação Automática Semissupervisionada de Papéis Semânticos para o Português do Brasil." Dissertação de Mestrado, Universidade de São Paulo.

Baptista, Jorge, Caroline Hagège, and Nuno Mamede. 2008. "Identificação, Classificação e Normalização de Expressões Temporais do Português: a Experiência do Segundo HAREM e o Futuro." In *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: o Segundo HAREM,* edited by Cristina Mota and Diana Santos, 33-54. Linguateca.

Bick, Eckhard. 2000. *The Parsing System Palavras. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus: Aarhus University Press.

Breiman, Leo, Jerome Friedman, R.A. Olshen, and Charles Stone. 1984. *Classification and Regression Trees*. New York: Chapman and Hall.

Cardoso, Nuno. 2008. "REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto." In *Desafios na Avaliação Conjunta do*

Reconhecimento de Entidades Mencionadas: O Segundo HAREM, edited by Cristina Mota and Diana Santos, 195-211. Linguateca.

Cardoso, C. Paula, Erick G. Maziero, Maria C. Jorge, Eloise M. Seno, Ariani Di Felippo, Lucia H. Rino Maria das Graças V. Nunes, and Thiago A. S. Pardo. 2011. "CSTNews - A Discourse Annotated Corpus for Single and Multi-document Summarization of News Texts in Brazilian Portuguese." Paper presented at the 3rd RST Brazilian Meeting, 88-105. Cuiabá, Mato Grosso, Brazil, October 24-26.

Genest, Pierre-Etienne, Guy Lapalme, and Mehdi Yousfi-Monod. 2009. "Hextac: the Creation of a Manual Extractive Run." Paper presented at the Second Text Analysis Conference, 1-6. Gaithersburg, Maryland, USA, November 14-15.

Genest, Pierre-Etienne, and Guy Lapalme. 2012. "Fully Abstractive Approach to Guided Summarization." Paper presented at the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, 2:354-358. Stroudsburg, Pennsylvania, USA.

Grishman, Ralph, and Beth Sundheim. 1996. "Message Understanding Conference - 6: a Brief History." Paper presented at the 16th International Conference on Computational Linguistics, COLING'96, 39:446-471. Stroudsburg, Pennsylvania, USA.

Haykin, Simon. 1998. *Neural Networks: A Comprehensive Foundation*. New Jersey: Prentice Hall - PTR. Second edition.

Karlsson, Fred. 1990. "Constraint Grammar as a Framework for Parsing Running Text." Paper presented at the 13th International Conference on Computational Linguistics, COLING'90, 3:168-173. Stroudsburg , Pennsylvania, USA.

Li, Peng, Yinglin Wang, Wei Gao, and Jing Jiang. 2011. "Generating Aspect-oriented Multi-document Summarization with Event-aspect Model." Paper presented at the Conference on Empirical Methods in Natural Language Processing, EMNLP'11, 1137-1146. Stroudsburg, Pennsylvania, USA.

Mann, Wiliam, and Sandra Thompson. 1987. *Rhetorical Structure Theory: A Theory of Text Organization*, Reprinted from the Structure of Discourse, ISI Reprint Series, 87-190. California: University of Southern California.

Mitchell, Tom. 1997. *Machine learning*. New York: McGraw-Hill.

Nadeau, David, and Satoshi Sekine. 1998. "A Survey of Named Entity Recognition and Classification." *Linguisticae Investigationes* 30:3-26.

Blake, C., and Christopher Merz. 1998. "UCI Repository of Machine Learning Databases." *Department of Information and Computer Science*.

Owczarzak, Karolina, and Hoa Dang. 2011. "Who Wrote What Where: Analyzing the Content of Human and Automatic Summaries." Paper presented at the Workshop on Automatic Summarization for Different Genres, Media, and Languages, 25-32. Portland, Oregon, USA, June 23.

Palmer, Martha, Daniel Gildea, and Nianguen Xue. 2010. “Semantic Role Labeling.” *Synthesis Lectures on Human Language Technologies* 3:1-103.

Radev, Dragomir. 2000. “A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-document Structure.” Paper presented at the 1st SIGdial Workshop on Discourse and Dialogue, SIGDIAL’00, 10:74-83. Stroudsburg, Pennsylvania, USA.

Rassi, Amanda P., Andressa C. Zacarias, Erick G. Maziero, Jackson W. Souza, Márcio S. Dias, Maria C. Jorge, Paula C. Cardoso, Pedro F. Balage, Renata T. Camargo, Verônica Agostini, Ariani Di Felippo, Eloise R. Seno, Lucia H. Rino, and Thiago A. S. Pardo. 2013. “Anotação de Aspectos Textuais em Sumários do Córpus CSTNews.” Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, NILC-TR-13-01, 394:1-59. São Carlos, São Paulo, São Paulo, Brasil, October.

Santos, Diana, and Nuno Cardoso. 2007. *Reconhecimento de Entidades Mencionadas em Português: Documentação e Actas do HAREM, a Primeira Avaliação Conjunta na Área*. Oslo/Lisboa: Linguateca.

Shamsfard, Mehrnoush, and Maryam Mousavi. 2008. “Thematic Role Extraction Using Shallow Parsing.” *International Journal of Computational Intelligence* 2(6):695-701.

Steinberger, Josef, Hristo Tanev, Mijail Kabadjov, and Ralf Steinberger. 2011. “JRC's Participation in the Guided Summarization Task at TAC 2010.” Paper presented at the Third Text Analysis Conference, TAC’10, 1-12. Gaithersburg, Maryland, USA, November 15-16.

Swales, Jhon. 1999. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Teufel, Simone, and Marc Moens. 1999. “Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting.” *Advances in Automatic Text Summarization*, 155:1-171.

Teufel, Simone, and Marc Moens. 2002. “Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status.” *Computational Linguistics* 28(4):409-445.

Tsoumakas, Grigorios, and Ioannis Katakis. 2007. “Multi-label Classification: an Overview.” *International Journal on Data Warehousing and Mining* 3:1-13.

Vapnik, Vladimir. 2000. *The Nature of Statistical Learning Theory*. New York: Springer Science & Business Media.

White, Michael, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce, and Kiri Wagsta. 2001. “Multidocument Summarization Via Information Extraction.” Paper presented at the First International Conference on Human Language Technology Research, HLT’01, 1-7. Stroudsburg, Pennsylvania, USA.

Zhu, Xiaojin, and Andrew Goldberg. 2009. “Introduction to Semi-Supervised Learning.” *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3:1-130.

Apêndice A – Aspectos nas categorias do CSTNews

Macroaspectos	Microaspectos
COMMENT	WHO_AGENT
COMPARISON	WHO_AFFECTED
CONSEQUENCE	WHEN
COUNTERMEASURES	WHERE
DECLARATION	WHY
GOAL	HOW
HISTORY	
PREDICTION	
SITUATION	
WHAT	

Tabela 44: Aspectos identificados na categoria *Cotidiano*

Macroaspectos	Microaspectos
COMMENT	WHO_AGENT
COMPARISON	WHO_AFFECTED
CONSEQUENCE	WHEN
DECLARATION	WHERE
GOAL	WHY
HISTORY	SCORE
PREDICTION	SITUATION
WHAT	
HOW	

Tabela 45: Aspectos identificados na categoria *Esportes*

Macroaspectos	Microaspectos
CONSEQUENCE	WHO_AGENT
DECLARATION	WHO_AFFECTED
COUNTERMEASURES	WHEN
HISTORY	WHERE
PREDICTION	WHY
WHAT	GOAL
	SITUATION

Tabela 46: Aspectos identificados na categoria *Mundo*

Macroaspectos	Microaspectos
COUNTERMEASURES	WHO_AGENT
COMPARISON	WHO_AFFECTED
CONSEQUENCE	WHEN
DECLARATION	WHERE
GOAL	WHY
HISTORY	HOW
PREDICTION	
WHAT	
SITUATION	

Tabela 47: Aspectos identificados na categoria *Política*

Apêndice B – Regras criadas para identificação de microaspectos

PESSOA = [H, HH, Hattr, Hbio, Hfam, Hideo, Hmyth, Hnat, Hprof, Hsick, Htit, hum]

ORGANIZAÇÃO = [admin, org, inst, media, party, suborg]

⊂ = "está contido em"
⊄ = "não está contido em"

Regra 1: Se o segmento APS contiver algum *token* associado a uma etiqueta semântica do tipo PESSOA/ORGANIZAÇÃO[14], e se o *token* não pertencer ao "léxico_de_local" do REPENTINO[15], então o segmento será corretamente anotado como WHO_AGENT/WHO_AFFECTED.

Entrada: *"<aspect APS=WHO_AGENT>O presidente</aspect> diz que algumas de as obras já estão em andamento, <aspect APS="WHO_AGENT">outras</aspect> vão começar logo."*

presidente_(Hprof) ⊂ PESSOA ⊄ léxico_de_local
outras_(diff) ⊄ PESSOA

Saída: *"<aspect APS=WHO_AGENT>O presidente </aspect> diz que algumas de as obras já estão em andamento, outras vão começar logo."*

Entrada: *"<aspect APS=WHO_AGENT> Ao menos 549 pessoas </aspect> morreram em consequência das enchentes que atingiram <aspect APS=WHO_AGENT> a Coréia-do-Norte </aspect> em julho"*

pessoas_(H) ⊂ PESSOA ⊄ léxico_de_local
Coréia-do-Norte_(hum) ⊂ PESSOA ⊂ léxico_de_local

Saída: *"<aspect APS=WHO_AGENT>Ao menos pessoas</aspect> morreram em consequência das enchentes que atingiram a Coréia-do-Norte em julho."*

Figura 20: Regras do *microaspecto* WHO_AGENT/WHO_AFFECTED

[14] O sistema PALAVRAS fornece etiquetas semânticas para cada *token*/palavra da sentença. Neste caso, só foram escolhidas as etiquetas cujas categorias representam entidades do tipo PESSOA e ORGANIZAÇÃO: http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html#semtags_nouns

[15] O REPENTINO (**REP**ositório para reconhecimento de **ENTI**dades com **NO**me) é um léxico desenvolvido no âmbito da participação no HAREM-2005. O REPENTINO está organizado com as seguintes categorias principais: "entidades abstratas", "arte/média/comunicação", "natureza", "eventos", "material impresso", "locais", "seres", etc. Neste caso, só foi usado o léxico de locais: http://www.linguateca.pt/repentino/

PREP = [de, em, a, por, para]

PRON = [ele(s), ela(s), este(s), esta(s), esse(s), essa(s), aquele(s), isto, isso, aquilo, aqui, aí, ali, outro(s), outra(s)]

ARTG = [a(s), o(s), um, uns, uma, umas, à(s)]

dia_da_semana = [segunda-feira, terça-feira, quarta-feira, quinta-feira, sexta-feira, sábado, domingo]

adverbio_de_tempo = [hoje, amanhã, ontem, anteontem, tarde, madrugada, noite, meia-noite, manhã]

lexico_de_tempo = [microssegundo, segundo, minuto, hora, dia, semana, mês, ano, década, milênio, semestre, bimestre, trimestre, época, tempo]

"+/-" = seguido_ou_não

Regra 1: Se a sentença tiver PREP + (PRON | ARTG) + averbio_de_tempo + PREP + (PRON | ARTG) + dia_da_semana +/- NÚMERO, então a sentença será anotada como WHEN.

"A chuva complicava o trânsito na manhã desta segunda-feira, 16."
na_(PREP+PRON) + manhã_(adverbio_de_tempo) + desta_(PREP+PRON) + segunda-feira_(dia_da_semana) + NÚMERO

"Uma nova série de ataques criminosos foi registrada na madrugada desta terça-feira.
na_(PREP+PRON) + madrugada_(adverbio_de_tempo) + desta_(PREP+PRON) + terça-feira_(dia_da_semana)

Regra 2: Se a sentença tiver PREP + (PRON | ARTG) + dia_da_semana, então a sentença será anotada como WHEN.

"Um terremoto atingiu Japão nesta segunda-feira matando 9 pessoas."
nesta_(PREP+PRON) + segunda-feira_(dia_da_semana)

Regra 3: Se a sentença tiver PREP + (PRON | ARTG) +/- (TOKEN | NÚMERO) + lexico_de_tempo, então a sentença será anotada como WHEN.

"Aos 18 minutos, Maicon fez o primeiro gol."
Aos_(PREP+ARTG) + 18_(NÚMERO) + minutos_(lexico_de_tempo)

"No primeiro tempo houve outras jogadas..."
No_(PREP +ARTG) + primeiro_(TOKEN) + tempo_(lexico_de_tempo)

"Os acontecimentos ocorreram nessa semana."
nessa__(PREP+PRON) semana_(lexico_de_tempo)

Regra 4: Se a sentença tiver PREP + (PRON | ARTG) + avérbio_de_tempo, então a sentença será anotada como WHEN.

"A quarta medida foi aprovada nesta madrugada."
nessa_(PREP+PRON) madrugada_(avérbio_de_tempo)

Figura 21: Regras do *microaspecto* WHEN

Regra 1: Se o segmento APS contiver a PREPOSIÇÃO “em”, seguida ou não de ARTIGO/PRONOME, seguido de um SUBSTANTIVO que não seja uma “expressão de tempo” ou um “advérbio de modo”, então o segmento será corretamente anotado como WHERE.

Entrada: *"Eu guardei as informações<aspect APS=WHERE>nesse computador</aspect>."*

em_(PREPOSIÇÃO) + esse_(PRONOME) + computador_(SUBSTANTIVO)

Saída: *"Eu guardei as informações<aspect APS=WHERE>nesse computador</aspect>."*

Entrada: *"<aspect APS=WHERE>No domingo</aspect>, uma batalha sangrenta ocorreu."*

em_(PREPOSIÇÃO) + o_(ARTIGO) + domingo _(expressão_de_tempo)

Saída: *“No domingo, uma batalha sangrenta ocorreu.”*

Entrada: *"Eu pense <aspect APS=WHERE>em voz alta</aspect>.”*

em_(PREPOSIÇÃO) + voz alta_(adverbo_de_modo)

Saída: *"Eu pense em voz alta.”*

Regra 2: Se a sentença tiver a PREPOSIÇÃO “em” + expressão_capitalizada, então a sentença será anotada como WHERE.

“O senador Marcos nasceu em São Paulo”
em_(PREPOSIÇÃO) + São Paulo_(expressão_capitalizado)

Figura 22: Regras do *microaspecto* WHERE

léxico_de_causa = [por isso, com isso, porque, devido a, por causa de, por força de, em função de, em virtude de, em razão de, em decorrência de, em consequência de, pois, visto que, já que, causado, prejudicado]

Regra 1: Se a sentença tiver expressão "léxico_de causa", então a sentença será anotada como WHY.

"O senador teve seu estado de saúde piorado, por causa de complicações gastrointestinais.”
por causa de_(léxico_de_causa)

“O encontro previsto para a noite foi cancelado porque os tucanos entenderam que ...”
porque_(léxico_de_causa)

Regra 2: Se a sentença tiver PREPOSIÇÃO "por" + verbo_infinitivo, então a sentença será anotada como WHY.

“Já Poliana Okimoto ficará fora de a decisão de os 800m livre por estar com infecção intestinal.”
por_(PREPOSIÇÃO) + estar_(verbo_infinitivo)

Regra 3: Se a sentença tiver a expressão "graças a" + ARTIGO, sem ser parte da expressão "dar graças a", então a sentença será anotada com aspecto WHY.

"Graças ao médico, o paciente não morreu."
graças a_(expressão) + o_(ARTIGO)

Figura 23: Regras do *microaspecto* WHY

léxico_de_score = [set(s), gol(s), jogo(s)]

Regra 1: Se a sentença tiver NÚMERO + léxico_de_score + "a" + NÚMERO, então a sentença será anotada como SCORE.

"A equipe brasileira venceu a Finlândia por 3 sets a 0 na Liga Mundial de Vôlei-06."
3_(NUM) + sets_(léxico_de_score) + a + 0_(NÚMERO)

Regra 2: Se a sentença tiver NÚMERO + metros + NÚMERO, então a sentença será anotada como SCORE.

"A medalha de prata ficou com a americana April Steiner com 4m40 e a de bronze com a cubana Yarisley Silva com 4m30."
4_(NÚMERO) + m_(metros) + 40_(NÚMERO)
4_(NÚMERO) + m_(metros) + 30_(NÚMERO)

Regra 3: Se a sentença tiver NÚMERO + minuto + NÚMERO + segundo + NÚMERO, então a sentença será anotada como SCORE.

"Eles fizeram história a o cravar o tempo de 7min12s27 e superar os Estados Unidos."
7_(NÚMERO) + min_(minuto) + 12_(NÚMERO) + s_(segundo) + 27_(NÚMERO)

"O Brasil conquistou a medalha de ouro na prova de natação, com o tempo de 3min15s90."
7_(NÚMERO) + min_(minuto) + 12_(NÚMERO) + s_(segundo) + 90_(NÚMERO)

Figura 24: Regras do *microaspecto* SCORE

Apêndice C – Resultados dos classificadores

Microaspecto	Classificador	Cobertura	Precisão	F1	Acurácia
WHO_AGENT	(1, 1) bag_of_words	0.436+/-0.12	0.81+/-0.12	0.567+/-0.12	0.732+/-0.06
	(2, 2) bag_of_words	0.051+/-0.06	1.0+/-0.76	0.098+/-0.1	0.619+/-0.02
	(2, 3) bag_of_words	0.051+/-0.06	1.0+/-0.84	0.098+/-0.11	0.619+/-0.02
	(1, 1) lemmas	0.462+/-0.15	0.818+/-0.14	0.59+/-0.12	0.742+/-0.06
	(2, 2) lemmas	0.051+/-0.06	1.0+/-0.76	0.098+/-0.1	0.619+/-0.02
	(2, 3) lemmas	0.051+/-0.06	1.0+/-0.84	0.098+/-0.11	0.619+/-0.02
	(1, 1) POS	0.487+/-0.10	0.655+/-0.1	0.559+/-0.09	0.691+/-0.05
	(2, 2) POS	0.487+/-0.13	0.613+/-0.11	0.543+/-0.09	0.67+/-0.07
	(2, 3) POS	0.462+/-0.14	0.581+/-0.14	0.514+/-0.13	0.649+/-0.09
	(1, 1) semantic	0.513+/-0.19	0.606+/-0.1	0.556+/-0.13	0.67+/-0.07
	(2, 2) semantic	0.436+/-0.12	0.68+/-0.16	0.531+/-0.07	0.691+/-0.05
	(2, 3) semantic	0.436+/-0.16	0.654+/-0.12	0.523+/-0.13	0.68+/-0.06
	(1, 1) lemmas+POS	0.410+/-0.18	0.8+/-0.17	0.542+/-0.14	0.722+/-0.06
	(2, 2) lemmas+POS	0.051+/-0.06	1.0+/-0.76	0.098+/-0.1	0.619+/-0.02
	(2, 3) lemmas+POS	0.051+/-0.06	1.0+/-0.84	0.098+/-0.11	0.619+/-0.02
	(1, 1) POS+semantic	0.462+/-0.13	0.621+/-0.09	0.529+/-0.11	0.67+/-0.06
	(2, 2) POS+semantic	0.487+/-0.14	0.633+/-0.1	0.551+/-0.1	0.68+/-0.06
	(2, 3) POS+semantic	0.538+/-0.15	0.636+/-0.08	0.583+/-0.11	0.691+/-0.06
WHO_AFFECTED	(1, 1) bag_of_words	0.167+/-0.1	0.75+/-0.4	0.273+/-0.16	0.835+/-0.03
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(1, 1) lemmas	0.222+/-0.16	1.0+/-0.3	0.364+/-0.22	0.854+/-0.03
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(1, 1) POS	0.056+/-0.07	0.5+/-0.48	0.1+/-0.12	0.814+/-0.01
	(2, 2) POS	0.111+/-0.11	0.333+/-0.35	0.167+/-0.16	0.794+/-0.04
	(2, 3) POS	0.222+/-0.17	0.4+/-0.23	0.286+/-0.17	0.794+/-0.05
	(1, 1) semantic	0.278+/-0.22	0.5+/-0.15	0.357+/-0.18	0.814+/-0.04
	(2, 2) semantic	0.111+/-0.12	0.5+/-0.4	0.182+/-0.17	0.814+/-0.03
	(2, 3) semantic	0.111+/-0.18	0.4+/-0.51	0.174+/-0.26	0.804+/-0.05
	(1, 1) lemmas+POS	0.222+/-0.13	0.8+/-0.34	0.348+/-0.18	0.845+/-0.03
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.814+/-0.0
	(1, 1) POS+semantic	0.222+/-0.14	0.5+/-0.27	0.308+/-0.18	0.814+/-0.05
	(2, 2) POS+semantic	0.222+/-0.17	0.571+/-0.25	0.32+/-0.2	0.825+/-0.04
	(2, 3) POS+semantic	0.222+/-0.16	0.5+/-0.29	0.308+/-0.19	0.812+/-0.05
WHEN	(1, 1) bag_of_words	0.091+/-0.11	0.667+/-0.6	0.16+/-0.17	0.781+/-0.02
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.763+/-0.01
	(2, 3) bag_of_words	0.0+/-0.03	0.0+/-0.4	0.0+/-0.06	0.763+/-0.01
	(1, 1) lemmas	0.174+/-0.14	0.667+/-0.39	0.276+/-0.21	0.784+/-0.04
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.763+/-0.01
	(2, 3) lemmas	0.0+/-0.03	0.0+/-0.4	0.0+/-0.06	0.763+/-0.01
	(1, 1) POS	0.043+/-0.09	0.5+/-0.75	0.08+/-0.15	0.763+/-0.04
	(2, 2) POS	0.13+/-0.15	0.375+/-0.27	0.194+/-0.17	0.742+/-0.05
	(2, 3) POS	0.261+/-0.21	0.375+/-0.17	0.308+/-0.18	0.722+/-0.06
	(1, 1) semantic	0.522+/-0.13	0.75+/-0.21	0.615+/-0.14	0.845+/-0.06
	(2, 2) semantic	0.391+/-0.11	0.75+/-0.18	0.514+/-0.12	0.825+/-0.04
	(2, 3) semantic	0.435+/-0.1	0.769+/-0.14	0.556+/-0.1	0.835+/-0.03
	(1, 1) lemmas+POS	0.174+/-0.14	0.667+/-0.45	0.276+/-0.21	0.784+/-0.06
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.763+/-0.01
	(2, 3) lemmas+POS	0.0+/-0.03	0.0+/-0.4	0.0+/-0.06	0.763+/-0.01
	(1, 1) POS+semantic	0.348+/-0.25	0.727+/-0.2	0.471+/-0.24	0.814+/-0.06
	(2, 2) POS+semantic	0.478+/-0.06	0.688+/-0.18	0.564+/-0.08	0.825+/-0.04

	(2, 3) POS+semantic	0.478+/-0.16	0.611+/-0.13	0.537+/-0.14	0.804+/-0.05
WHERE	(1, 1) bag_of_words	0.118+/-0.13	1.0+/-0.0	0.211+/-0.19	0.845+/-0.02
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(1, 1) lemmas	0.176+/-0.17	0.75+/-0.6	0.286+/-0.26	0.845+/-0.03
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(1, 1) POS	0.059+/-0.17	0.25+/-0.34	0.095+/-0.22	0.804+/-0.02
	(2, 2) POS	0.294+/-0.15	0.455+/-0.29	0.357+/-0.14	0.814+/-0.06
	(2, 3) POS	0.353+/-0.16	0.5+/-0.27	0.414+/-0.17	0.825+/-0.07
	(1, 1) semantic	0.412+/-0.16	0.636+/-0.16	0.5+/-0.15	0.856+/-0.03
	(2, 2) semantic	0.235+/-0.21	0.667+/-0.42	0.348+/-0.27	0.845+/-0.05
	(2, 3) semantic	0.235+/-0.17	0.667+/-0.21	0.348+/-0.2	0.845+/-0.03
	(1, 1) lemmas+POS	0.118+/-0.17	0.667+/-0.6	0.2+/-0.26	0.835+/-0.03
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.825+/-0.0
	(1, 1) POS+semantic	0.353+/-0.23	0.6+/-0.32	0.444+/-0.23	0.845+/-0.05
	(2, 2) POS+semantic	0.412+/-0.34	0.636+/-0.21	0.5+/-0.31	0.856+/-0.06
	(2, 3) POS+semantic	0.471+/-0.22	0.615+/-0.22	0.533+/-0.19	0.856+/-0.05
WHY	(1, 1) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.01
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(1, 1) lemmas	0.0+/-0.06	0.0+/-0.6	0.0+/-0.11	0.897+/-0.01
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(2, 2) POS	0.1+/-0.13	0.333+/-0.62	0.154+/-0.19	0.887+/-0.03
	(2, 3) POS	0.2+/-0.23	0.333+/-0.33	0.25+/-0.24	0.876+/-0.03
	(1, 1) semantic	0.1+/-0.13	0.333+/-0.47	0.154+/-0.2	0.887+/-0.03
	(2, 2) semantic	0.0+/-0.13	0.0+/-0.83	0.0+/-0.22	0.897+/-0.01
	(2, 3) semantic	0.0+/-0.1	0.0+/-0.79	0.0+/-0.17	0.897+/-0.01
	(1, 1) lemmas+POS	0.0+/-0.09	0.0+/-0.92	0.0+/-0.17	0.897+/-0.02
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.897+/-0.0
	(1, 1) POS+semantic	0.1+/-0.13	0.5+/-0.74	0.167+/-0.21	0.897+/-0.03
	(2, 2) POS+semantic	0.1+/-0.2	0.333+/-0.62	0.154+/-0.3	0.887+/-0.04
	(2, 3) POS+semantic	0.2+/-0.2	0.5+/-0.56	0.286+/-0.28	0.897+/-0.04
HOW	(1, 1) bag_of_words	0.25+/-0.23	1.0+/-0.8	0.4+/-0.35	0.938+/-0.02
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(1, 1) lemmas	0.25+/-0.31	1.0+/-0.6	0.4+/-0.4	0.938+/-0.03
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(2, 2) POS	0.0+/-0.12	0.0+/-0.6	0.0+/-0.18	0.907+/-0.02
	(2, 3) POS	0.125+/-0.18	0.25+/-0.54	0.167+/-0.2	0.897+/-0.04
	(1, 1) semantic	0.25+/-0.35	1.0+/-0.79	0.4+/-0.44	0.938+/-0.02
	(2, 2) semantic	0.0+/-0.12	0.0+/-0.98	0.0+/-0.22	0.918+/-0.01
	(2, 3) semantic	0.125+/-0.16	1.0+/-0.81	0.222+/-0.25	0.928+/-0.01
	(1, 1) lemmas+POS	0.25+/-0.35	1.0+/-0.8	0.4+/-0.47	0.938+/-0.03
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.918+/-0.0
	(1, 1) POS+semantic	0.0+/-0.12	0.0+/-1.0	0.0+/-0.22	0.918+/-0.02
	(2, 2) POS+semantic	0.0+/-0.11	0.0+/-0.34	0.0+/-0.17	0.907+/-0.02
	(2, 3) POS+semantic	0.125+/-0.17	0.5+/-0.77	0.2+/-0.23	0.918+/-0.02
SITUATION	(1, 1) bag_of_words	0.333+/-0.28	1.0+/-0.63	0.5+/-0.36	0.959+/-0.02
	(2, 2) bag_of_words	0.167+/-0.21	1.0+/-0.8	0.286+/-0.32	0.948+/-0.01

	(2, 3) bag_of_words	0.167+/-0.21	1.0+/-0.8	0.286+/-0.32	0.948+/-0.01
	(1, 1) lemmas	0.333+/-0.31	0.667+/-0.64	0.444+/-0.37	0.948+/-0.02
	(2, 2) lemmas	0.167+/-0.2	1.0+/-0.92	0.286+/-0.32	0.948+/-0.02
	(2, 3) lemmas	0.167+/-0.2	1.0+/-0.92	0.286+/-0.32	0.948+/-0.02
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.938+/-0.01
	(2, 2) POS	0.0+/-0.13	0.0+/-0.8	0.0+/-0.23	0.938+/-0.02
	(2, 3) POS	0.333+/-0.27	0.5+/-0.6	0.4+/-0.27	0.938+/-0.02
	(1, 1) semantic	0.333+/-0.26	0.5+/-0.49	0.4+/-0.3	0.938+/-0.03
	(2, 2) semantic	0.167+/-0.3	1.0+/-0.79	0.286+/-0.39	0.948+/-0.02
	(2, 3) semantic	0.167+/-0.31	1.0+/-0.76	0.286+/-0.39	0.948+/-0.02
	(1, 1) lemmas+POS	0.333+/-0.34	1.0+/-0.74	0.5+/-0.44	0.958+/-0.02
	(2, 2) lemmas+POS	0.167+/-0.2	1.0+/-0.92	0.286+/-0.32	0.948+/-0.02
	(2, 3) lemmas+POS	0.167+/-0.2	1.0+/-0.92	0.286+/-0.32	0.948+/-0.02
	(1, 1) POS+semantic	0.333+/-0.31	0.667+/-0.54	0.444+/-0.35	0.948+/-0.03
	(2, 2) POS+semantic	0.333+/-0.33	0.667+/-0.58	0.444+/-0.35	0.948+/-0.02
	(2, 3) POS+semantic	0.333+/-0.29	0.667+/-0.6	0.444+/-0.34	0.948+/-0.02
SCORE	(1, 1) bag_of_words	0.0+/-0.2	0.0+/-0.6	0.0+/-0.3	0.969+/-0.01
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(1, 1) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.01
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(2, 2) POS	0.0+/-0.2	0.0+/-0.3	0.0+/-0.24	0.969+/-0.01
	(2, 3) POS	0.0+/-0.33	0.0+/-0.78	0.0+/-0.45	0.969+/-0.02
	(1, 1) semantic	0.0+/-0.45	0.0+/-0.67	0.0+/-0.48	0.969+/-0.02
	(2, 2) semantic	0.0+/-0.27	0.0+/-0.8	0.0+/-0.4	0.969+/-0.01
	(2, 3) semantic	0.0+/-0.27	0.0+/-0.8	0.0+/-0.4	0.969+/-0.01
	(1, 1) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.01
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.0
	(1, 1) POS+semantic	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.969+/-0.01
	(2, 2) POS+semantic	0.0+/-0.31	0.0+/-0.81	0.0+/-0.43	0.969+/-0.02
	(2, 3) POS+semantic	0.0+/-0.27	0.0+/-0.4	0.0+/-0.32	0.969+/-0.01

Tabela 48: Resultados dos classificadores usando AM