UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

IDENTIFICAÇÃO AUTOMÁTICA DE MACROASPECTOS EM TEXTOS JORNALÍSTICOS

ALESSANDRO YOVAN BOKAN GARAY
THIAGO ALEXANDRE SALGUEIRO PARDO

Nº 407

RELATÓRIOS TÉCNICOS

São Carlos – SP
Jul./2015

Instituto de Ciências Matemáticas e de Computação – ICMC-USP

Identificação Automática de Macroaspectos em Textos Jornalísticos

Alessandro Yovan Bokan Garay
Thiago Alexandre Salgueiro Pardo

NILC-TR-15-02

Julho, 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 *macroaspectos* utilizando técnicas de aprendizado de máquina e regras manuais. A metodologia foi avaliada sobre o córpus de notícias jornalísticas CSTNews, previamente anotado manualmente com aspectos informativos. Os resultados são satisfatórios usando regras manuais, demostrando que alguns *macroaspectos* 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: "o que aconteceu", "quando aconteceu", "onde aconteceu", "quais foram as contramedidas".

Os aspectos foram propostos 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 naturais ameaçados" e "Julgamentos e investigações". 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 Figura 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 (por exemplo, Steinberger et al., 2010; Li et al., 2011; Genest e Lapalme, 2012). Porém, o uso de

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

[2] Terminologia em inglês proposta pela TAC.

aspectos não é novidade em sumarização e nem em outras áreas da Linguística e do Processamento de Linguagem Natural (PLN). Por exemplo, Swales (1999) propõe o uso de aspectos como componentes 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 Textos, 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 automaticamente aspectos (*macroaspectos*, no caso deste relatório) é **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), dando-se continuidade ao trabalho relatado por Bokan e Pardo (2015). Enquanto neste trabalho anterior o foco estava na identificação de microaspectos, neste relatório foca-se nos macroaspectos, conforme será detalhado posteriormente.

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. *Text Analysis Conference* (TAC)

A TAC é a principal conferência e competição científica dedicada à sumarização automática, fornecendo uma grande coleção de dados de teste, procedimentos de avaliação e um fórum para compartilhar resultados. No ano 2010[3], a TAC propôs a tarefa de Sumarização Guiada[4] (em inglês, *Guided Summarization*), com a finalidade de produzir um sumário de 100 palavras a partir de um conjunto de 10 artigos para um tópico dado. Todos os participantes tinham uma lista de aspectos para cada categoria textual, e cada sumário produzido deveria conter todos os aspectos designados para cada categoria. As categorias indicam o assunto ou domínio do texto (por exemplo: política, esporte, economia, etc.). Os aspectos podem ser específicos para

[3] http://www.nist.gov/tac/2010/Summarization/Guided-Summ.2010.guidelines.html
[4] Modalidade da SA multidocumento "assistida" por aspectos informativos, que visa construir sumários orientados pelo significado.

cada categoria ou genéricos, indicando sua validade ou abrangência para um leque maior de domínios. As categorias e seus aspectos definidos pela TAC incluem:

- **Acidentes e desastres naturais:** descrição do fato (WHAT), data (WHEN), localização (WHERE), razões do acidente/desastre (WHY), entidade afetada (WHO_AFFECTED), danos (DAMAGES), esforços de resgate/contramedidas (COUNTERMEASURES).
- **Ataques:** descrição do fato (WHAT), data (WHEN), localização (WHERE), entidade afetada (WHO_AFFECTED), danos (DAMAGES), criminosos (PERPETRATORS), esforços de resgate/contramedidas (COUNTERMEASURES).
- **Saúde e segurança:** qual é o problema (WHAT), quem foi afetado (WHO_AFFECTED), como foi afetado (HOW), por que isso acontece (WHY), contramedidas (COUNTERMEASURES).
- **Recursos naturais ameaçados:** descrição do recurso (WHAT), importância do recurso (IMPORTANCE), ameaças (THREATS), contramedidas (COUNTERMEASURES).
- **Julgamentos e investigações:** quem está sob investigação (WHO), quem está investigando ou processando (WHO_INV), o por quê (WHY), acusações específicas (CHARGES), sentença/consequência (SENTENCE), como é que se reagiu às acusações (PLEAD).

Muitos estudos foram desenvolvidos seguindo esses princípios da TAC 2010. Por exemplo, Steinberger et al. (2010) realizaram análises semânticas profundas para a modelagem de aspectos visando a SA multilíngue. Makino et al. (2012) e Li et al. (2011) compilaram aspectos de sumários da Wikipédia. Barrera et al. (2011) criaram um sistema de perguntas e respostas com base na identificação de aspectos para diferentes categorias. Mesmo antes da TAC, alguns trabalhos já apresentavam abordagens semelhantes, por exemplo, White et al. (2001) propuseram *templates* com base em aspectos para sumários de textos de desastres, e Zhou et al. (2005) estudaram os aspectos presentes em sumários biográficos.

Como já foi dito, os aspectos informativos podem ser dependentes de categoria, ou seja, podem ser específicos para cada categoria em particular. Por exemplo, os aspectos WHY e DAMAGES da categoria "Acidentes e desastres naturais" são diferentes dos aspectos IMPORTANCE, THREATS e COUNTERMEASURES da categoria "Recursos naturais ameaçados". Por outro lado, o aspecto WHAT é geral e se aplica a quase todas as categorias (da mesma forma ocorre com WHEN e WHERE para as categorias "Acidentes e desastres naturais" e "Ataques"). Por esse motivo, os aspectos devem ser devidamente analisados e definidos para cada categoria de um dado córpus. Na seguinte seção, narra-se o processo de anotação de aspectos sobre o córpus jornalístico em língua portuguesa CSTNews. Nessa anotação, mantiveram-se alguns aspectos propostos pela TAC e foram também definidos novos aspectos.

2.2. O córpus CSTNews

O córpus CSTNews[5] (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

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

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 Figura 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*.

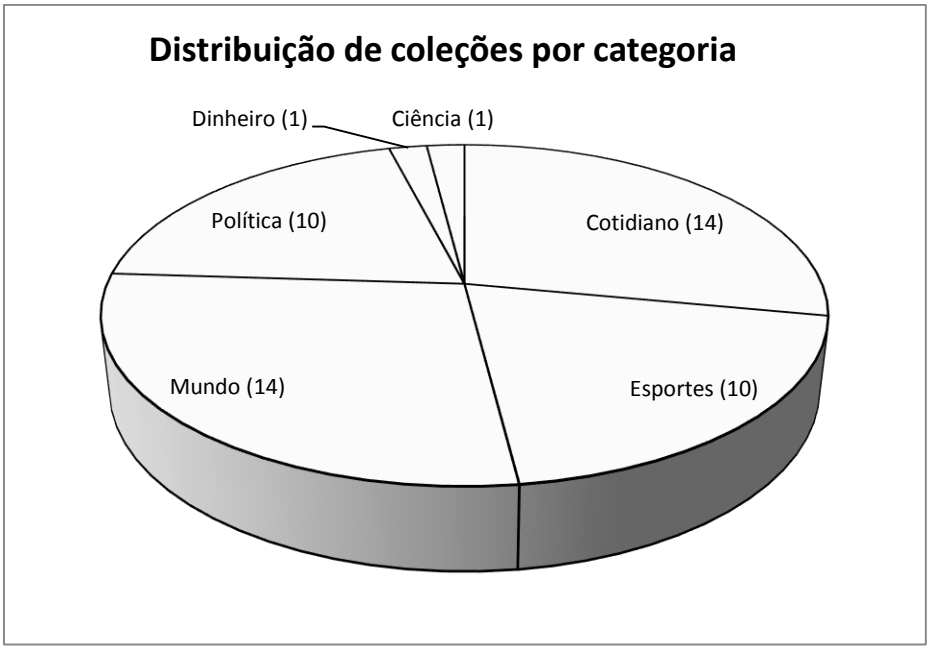

Figura 2: Distribuição das coleções por categoria

Além dos textos-fonte crus, o córpus CSTNews possui sumários manuais abstrativos monodocumento, sumários manuais abstrativos multidocumento e 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 50 sumários manuais multidocumento do córpus CSTNews.

2.2.1. 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 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 originalmente na TAC 2010. Contudo, existem similaridades com as seis categorias consideradas (ver Figura 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 compostos 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, realizou-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 (ver Tabela 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 *micro* e *macroaspectos*. Os *microaspectos* representam segmentos locais que compõem uma sentença. Os *macroaspectos* dependem do conteúdo sentencial em contexto. No total, foram identificados 11 *macroaspectos* e 9 *microaspectos*, apesar de haver alguma variação nesses conjuntos em função da categoria anotada (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 Figura 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).

[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

Por último, na anotação, foi relevante distinguir aspectos que transmitem informações principais daqueles relativos a informações secundárias. Diante disso, aos aspectos podia ser adicionado o sufixo EXTRA. Por exemplo, uma sentença é anotada como WHERE_EXTRA se possuir alguma informação de localidade que não se refere ao evento principal. Neste trabalho, não existe uma distinção entre ideias principais e secundárias. Portanto, os sufixos EXTRA foram ignorados, deixando os aspectos em suas formas originais (ver Tabela 1).

2.3. 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 Figura 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 indica 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 para fornecer informações léxicas, morfossintáticas e semânticas tanto para a criação do

classificador usando técnicas de Aprendizado de Máquina (ver Seção 3.1), quanto para a criação de regras manuais (ver Seção 3.2).

2.4. Papéis Retóricos

Os papeis retóricos indicam funções argumentativas e informativas dos segmentos textuais. Eles podem ser sinalizados por padrões linguísticos presentes na sentença. Assim, os *macroaspectos* são similares aos papeis retóricos por emergirem da informação contida nas sentenças em seus contextos.

Dayrel et al. (2012) propuseram um sistema que detecta padrões linguísticos particulares em *abstracts* de artigos científicos escritos em língua inglesa, denominado MAZEA (*Multi-label Argumentative Zoning for English Abstracts*). O sistema tenta identificar papéis retóricos, também chamados de zonas argumentativas, nas sentenças dos textos de gênero científico: BACKGROUNG (contexto), GAP (lacuna), PURPOSE (objetivo), METHOD (metodologia), RESULT (resultados) e CONCLUSION (conclusão). Devido ao fato de uma sentença poder ser anotada com mais de uma zona argumentativa, o problema de classificação torna-se multirrótulo. Os algoritmos de classificação resultaram da combinação dos algoritmos das bibliotecas Mulan[6], tais como *Classifier Chain* (Read et al., 2009) e *Rakel* (Tsoumakas e Vlahavas, 2007), e da biblioteca WEKA[7], como *Sequential Minimal Optimization* (SMO) (Platt, 1998), otimização do *Support Vector Machine* (SVM) (Vapnik, 2000); e *J48*, implementação *open source* do algoritmo C4.5 (Quinlan, 1993). Os melhores resultados foram obtidos pela combinação dos classificadores *Chain* + SMO.

O sistema MAZEA está baseado no sistema AZEA (*Argumentative Zoning for English Abstracts*) (Genoves Jr. et al., 2007), em que foi criado um classificador binário para identificar papéis retóricos nas sentenças. Utilizaram-se os algoritmos *J48*, SMO e Naïve Bayes. Os melhores resultados foram obtidos pelo Naïve Bayes, seguido do SMO. Tanto o sistema MAZEA quanto o sistema AZEA baseiam-se na tentativa de identificar movimentos retóricos em textos científicos proposta por Teufel e Moens (2002) e Feltrim et al. (2006), adotando um enfoque linguístico profundo. Outros sistemas são independentes da categoria textual e usam um enfoque superficial com base na estatística somente.

A extração de atributos é um passo importante na hora de se criar um classificador de papéis retóricos. Teufel (1999) define um total de 12 tipos de atributos (ver Tabela 2). Tanto o MAZEA quanto o AZEA se baseiam na extração de 6 atributos na criação dos classificadores multirrótulo e binário, respectivamente: *tamanho*, *posição*, *tempo*, *voz*, *modal* e *expressão padrão*.

O atributo *TF-IDF* visa identificar termos significativos que são frequentes numa sentença, mas que são raros nas outras sentenças do documento. Tal atributo é bastante usado na tarefa de Recuperação de Informação (Salton e McGill, 1989). Assim, por exemplo, uma sentença que contenha termos específicos relacionados à metodologia de pesquisa pode representar o papel retórico METHOD. As palavras que ocorrem no *título* também são boas candidatas para

[6] http://sourceforge.net/projects/mulan/
[7] http://www.cs.waikato.ac.nz/ml/weka/

identificar papéis retóricos específicos no texto. Por exemplo, em PURPOSE, costuma-se colocar palavras relacionadas ao título. De maneira igual, as sentenças que contêm termos significativos ou palavras do título podem ser bons indicadores do *macroaspecto* WHAT.

O *tamanho* da sentença é um atributo trivial, já que não está relacionado diretamente aos papéis retóricos. Mas, mesmo assim, deve ser considerado porque indica a complexidade da sentença. A *posição* da sentença no texto e nos parágrafos é muito importante. Normalmente, as primeiras sentenças de um *abstract* científico descrevem BACKGROUND e PURPOSE, já as últimas descrevem RESULT e CONCLUSION. Da mesma forma, existem *macroaspectos* que sempre ocorrem na primeira sentença dos textos, como WHAT, WHEN e WHERE.

Atributo	Descrição	Valores
TF-IDF	A sentença contém termos significativos determinados pela medida TD-IDF?	Sim ou não
Título	A sentença contém palavras que ocorrem no título?	Sim ou não
Tamanho	A sentença possui um maior número de palavras que um limiar definido?	Sim ou não
Posição Texto	Posição da sentença no texto em relação a 10 segmentos	Faixas de A a J
Posição Parágrafo	Posição da sentença dentro de um paragrafo	Começo, meio e fim
Tempo	Tempo do primeiro verbo flexionado	9 tempos verbais
Voz	Voz do primeiro verbo flexionado	Ativa ou passiva
Modal	Presença de verbos auxiliares modais	Sim ou não
Citação 1	A sentença contém alguma citação ou nome de autor?	Citação, autor, nenhum
Citação 2	A sentença contém uma autocitação?	Sim, não, nenhum
Citação 3	Posição da citação na sentença	Começo, meio, fim, nenhum
Expressão Padrão	Primeira Expressão Padrão (EP) na sentença	20 tipos de EP

Tabela 2: Atributos definidos por Teufel (1999)

Os atributos *tempo*, *voz* e *modal* representam a morfossintaxe do verbo. Os textos científicos seguem um padrão de escrita bem definido, portanto, a análise do primeiro verbo da sentença é de muita importância. Por exemplo, nas sentenças do tipo BACKGROUND, o primeiro verbo costuma estar no tempo presente. Em textos jornalísticos, o padrão de escrita é diferente, sendo que todos os verbos da sentença possuem a mesma importância. Por exemplo, uma sentença pode ter um verbo no tempo pretérito perfeito, indicando HISTORY, e um verbo no tempo futuro, indicando PREDICTION.

As *citações* são referências que indicam o trabalho de outras pessoas. Tais trabalhos representam o estado da arte ou os trabalhos relacionados ao foco da pesquisa. Esse atributo é um claro identificador do papel retórico BACKGROUND. No gênero jornalístico, praticamente não existem esses tipos de citações, portanto, esse atributo não será útil para nosso objetivo.

Já as *expressões padrão* são combinações de palavras frequentes, as quais estão relacionadas aos papéis retóricos. Por exemplo, as expressões “acredita-se que” e “visa-se a" comumente são sinalizadoras de PURPOSE. De maneira igual, tais expressões estão relacionadas aos *macroaspectos*, por exemplo: “declarou que” e “antigamente” podem ser sinalizadoras de DECLARATION e HISTORY, respectivamente.

Como já foi dito anteriormente, os *macroaspectos* são parecidos com os papéis retóricos. Portanto, com base nos atributos da literatura (Teufel e Moens, 2002; Feltrim et al., 2006; Genoves Jr. et al., 2007; Dayrel et al., 2012), visa-se criar um classificador que identifique automaticamente *macroaspectos*. Cabe ressaltar que os textos a serem processados são do gênero jornalístico. Dessa forma, os atributos serão **adaptados** para o gênero em foco. Os atributos a serem utilizados são: *TF-IDF*, *título*, *posição*, *tamanho*, *tempo*, *voz*, *modal* e *expressão padrão*. Na seção seguinte, descreve-se a metodologia utilizada na identificação dos *macroaspectos*.

3. Metodologia

Neste relatório, o processo de identificação automática de *macroaspectos* foi dividido em 3 fases (ver Figura 5). 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 *macroaspectos* usando duas abordagens:
    a. **Aprendizado de Máquina (AM):** uso de técnicas de AM para criar um classificador de *macroaspectos*. Por um lado, serão criados classificadores com base nos atributos definidos por Teufel (1999) e utilizados em outros trabalhos da literatura (Teufel e Moens, 2002; Feltrim et al., 2006; Genoves Jr. et al., 2007; Dayrel et al., 2012). Por outro lado, serão criados classificadores com base em atributos léxico-semânticos: *bag of words*, lemas, *part-of-speech* (POS), etiquetas semânticas, e a combinação deles. Cabe ressaltar que a maioria dos atributos utilizados nos classificadores são fornecidos pelo *parser* PALAVRAS (Bick, 2000). Esta abordagem atende os *macroaspectos* WHAT, CONSEQUENCE, COMMENT, DECLARATION e HISTORY. O restante dos *macroaspectos* não foi considerado por haverem poucas sentenças anotadas.
    b. **Regras manuais:** devido ao baixo resultado obtido pela abordagem usando AM, optou-se pela criação de regras com base na identificação de padrões linguísticos presentes em todas as sentenças do córpus anotadas com *macroaspectos*. Esta abordagem atende os *macroaspectos* COMPARISON, DECLARATION, GOAL, HISTORY e PREDICTION. Não foi possível identificar padrões linguísticos para o restante dos *macroaspectos*.
3. Obter um conjunto de sentenças anotadas automaticamente com *macroaspectos*.

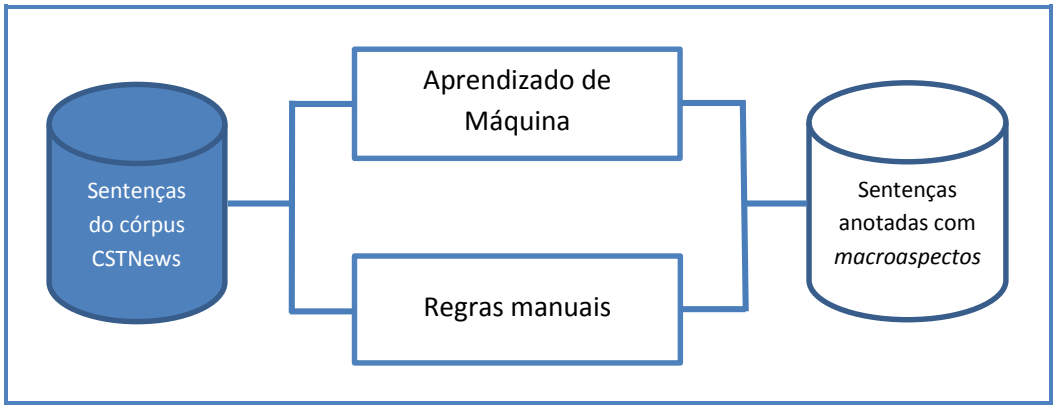

Figura 5: Metodologia do processo de identificação de *macroaspectos*

3.1. 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, teoricamente, representam conhecimento útil à 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 *macroaspectos*". 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.

A identificação de *macroaspectos* 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 os 5 melhores, para cada um dos *macroaspectos* WHAT, CONSEQUENCE, COMMENT, DECLARATION e HISTORY, respectivamente. Os *macroaspectos* COMPARISON, PREDICTION, COUNTERMEASURES, GOAL, SITUATION e HOW não foram considerados por terem **poucas** instâncias anotadas.

Por um lado, criaram-se classificadores binários com base nos atributos definidos por Teufel (1999) e utilizados em outros autores da literatura, como Teufel e Moens (2002), Feltrim et al. (2006), Genoves Jr. et al. (2007) e Dayrel et al. (2012) (ver Tabela 2). Como já foi dito, os atributos foram definidos originalmente para o gênero científico, portanto, os atributos foram adaptados para o gênero jornalístico.

Os atributos *TF-IDF, título*, *tamanho* e *posição* foram conservados na forma original. Já para os atributos *tempo*, *voz* e *modal,* não só foi considerado o primeiro verbo, mas sim todos os

verbos da sentença. Por exemplo, costuma-se classificar como PREDICTION as sentenças que possuem algum verbo no tempo futuro. No caso da sentença "Ele melhorou e está estável, mas continuará internado", o verbo "continuará" (tempo futuro) é o terceiro verbo da sentença. Isso acontece porque o estilo de escrita jornalístico não segue o mesmo padrão dos textos científicos. Finalmente, para o atributo *expressão padrão*, só foram identificadas expressões para DECLARATION (por exemplo, "de acordo com", "segundo" e os verbos ilocutórios) e COMPARISON ("em relação a", "em comparação a"). Cabe ressaltar que não foi possível identificar uma maior quantidade de expressões devido a pouca quantidade de sentenças anotadas no córpus CSTNews.

Por outro lado, criaram-se vários classificadores binários com base em 6 tipos de atributos léxico-semânticos (ver Tabela 3). Esses atributos foram utilizados no trabalho de Bokan e Pardo (2015), como parte do processo de identificação automática de *microaspectos*. Portanto, tais aspectos podem ser úteis, também, na identificação dos *macroaspectos.* Para extrair tais atributos, utilizou-se o formato *flat* (simples) do *parser* PALAVRAS (ver Figura 4). Cada atributo é representado por unigramas "(1, 1)", bigramas "(2, 2)" e bigramas + trigramas "(2, 3)". Assim, para cada um dos 5 *macroaspectos*, cria-se um classificador resultado da representação (unigramas, bigrama, bigrama+trigrama) de cada um dos 6 tipos de atributos. Por exemplo, o classificador denominado "*(2, 3) POS*" foi criado com base em todos os bigramas e trigramas "(2, 3)" das classes gramaticais (POS) de todas as palavras do córpus. No total, foram criados 90 classificadores binários (ver Apêndice C).

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 3: Atributos definidos

A técnica de aprendizado supervisionado utilizada, tanto com atributos de Teufel (1999) quanto com atributos léxico-semânticos, foi SVM. 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 o SVM, entre outras técnicas da literatura (SMO, Naïve Bayes, *J48*), por ser atualmente a técnica mais utilizada na literatura para classificação com textos. Além disso, é a melhor técnica em tratamento de vetores especiais de grandes dimensões.

Portanto, propõe-se o uso de AM para criar classificadores que possam identificar automaticamente *macroaspectos*. Objetiva-se obter o melhor classificador para cada

macroaspecto avaliando-se todos os possíveis classificadores gerados a partir dos atributos definidos por Teufel (1999) e os atributos léxico-semânticos. As instâncias de treino e teste são as sentenças dos sumários anotados do córpus CSTNews.

3.2. Regras manuais

No relatório da identificação dos *microaspectos* (Bokan e Pardo, 2015), foi proposto um sistema de anotação de microaspectos denominado "sistema APS" (Anotador de Papéis Semânticos). Para melhorar o desempenho de dito sistema, que até o momento já tinha conseguido resultados consideráveis, criaram-se regras manuais com base nos padrões linguísticos identificados nos "falsos negativos" e "falsos positivos". Os primeiros referem-se àquelas 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.

Nossa proposta da abordagem usando AM obteve resultados muito baixos, sendo que vários dos *macroaspectos* não puderam ser identificados corretamente em nenhuma sentença. Portanto, na criação de regras manuais para os macroaspectos, utilizaram-se **todas** as sentenças anotadas do córpus CSTNews com a finalidade de achar padrões linguísticos que pudessem representar alguns dos *macroaspectos*.

Foram criadas regras para os aspectos COMPARISON, DECLARATION, PREDICTION, HISTORY e GOAL. No entanto, não foram criadas regras para os aspectos WHAT, COMMENT, CONSEQUENCE, COUNTERMEASURES, SITUATION e HOW, por não terem sido achados padrões para criação de regras. No Apêndice B, apresenta-se, de maneira clara, o conjunto de regras definido para os *macroaspectos*.

A maioria das regras visam identificar *expressões padrão*. Assim, por exemplo, a expressão "em relação a" denota COMPARISON (ver Figura 14). De igual maneira, as expressões "segundo" e "de acordo com" correspondem a DECLARATION (ver Figura 15); "previsão" corresponde a PREDICTION (ver Figura 16); "desde" e "da história" correspondem a HISTORY (ver Figura 17); e "objetivo" corresponde a GOAL (ver Figura 18).

Outras regras se baseiam no tipo de verbo. Por exemplo, qualquer tipo de verbo ilocutório (declarar, afirmar, dizer, informar, anunciar, expressar, etc.) denota DECLARATION. Outras regras se baseiam no tempo verbal. Por exemplo, os verbos no futuro costumam expressar uma previsão ou PREDICTION.

Cabe ressaltar que a maioria das regras (COMPARISON, GOAL e PREDICTION) foram criadas sobre pouca quantidade de sentenças anotadas, sendo relativamente fácil de se identificar padrões linguísticos. No entanto, existe a possibilidade de acontecer *overfitting*[8] nas regras, por estas serem criadas e testadas sobre um conjunto mínimo de dados. Um modelo com *overfitting* apresenta uma alta precisão (ver Seção 4), porém tal modelo não é uma boa representação da realidade.

[8] Termo utilizado em AM ou estatística para dizer que o modelo estatístico se ajustou demasiadamente ao conjunto de dados, não sendo capaz de generalizar adequadamente.

4. Experimentos e resultados

As abordagens propostas (AM e regras manuais) foram avaliadas sobre um conjunto de sentenças anotadas manualmente com *macroaspectos*. Tais sentenças foram extraídas dos sumários multidocumento do córpus CSTNews (ver Seção 2.2.1). 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 Tabela 4. 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 4: Matriz de confusão

As estimativas de erro são calculadas por meio das quantidades de instâncias/exemplos, dando origem às métricas de avaliação. As métricas são calculadas conforme a classe positiva (SIM) e negativa (NÃO) para cada *macroaspecto*. 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 têm 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.

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

Na abordagem usando AM, os classificadores foram *treinados* e *testados* com as 322 sentenças (ou instâncias) do córpus CSTNews, anotadas com aspectos. A estratégia de treinamento e teste foi de repetidas divisões em subconjuntos (em várias iterações) com *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 (Newman et al., 1998). Portanto, o córpus foi 10 vezes estratificado aleatoriamente, 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). Não foi utilizado o tradicional *10-fold cross-validation* porque, devido ao córpus ser muito pequeno, o *fold* de teste teria poucas instâncias. Com a técnica de estratificação, garante-se uma melhor distribuição de classes e, por conseguinte, resultados mais justos.

Como já foi dito, a técnica de AM supervisionada usada foi 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"*** (225 instâncias de treino e 97 instâncias de teste). Esta abordagem atende os *macroaspectos* WHAT, CONSEQUENCE, COMMENT, DECLARATION e HISTORY. Os *macroaspectos* COMPARISON, PREDICTION, COUNTERMEASURES, GOAL, SITUATION e HOW não foram considerados por terem poucas instâncias anotadas.

Diferentemente da abordagem anterior, na abordagem usando regras manuais utilizaram-se as 322 sentenças do córpus para *teste*. Para avaliar esta abordagem, usaram-se as mesmas métricas de avaliação que na abordagem com AM. Como já foi dito, foram criadas regras manuais para os aspectos COMPARISON, DECLARATION, GOAL, HISTORY e PREDICTION, com base nos padrões identificados nas sentenças anotadas (ver Apêndice B). Os macroaspectos WHAT, COMMENT, CONSEQUENCE, COUNTERMEASURES, SITUATION e HOW não foram considerados.

Na Tabela 5, apresenta-se uma visão geral das abordagens utilizadas para cada *macroaspecto*. Observa-se que os aspectos COUNTERMEASURES, SITUATION e HOW não foram considerados, portanto, não serão apresentados resultados para esses aspectos. A seguir, apresentam-se os resultados obtidos pelas abordagens propostas para cada *macroaspecto*.

	Abordagens com AM		Abordagem com Regras
Macroaspecto	**Atributos de Teufel (1999)**	**Atributos léxico-semântico**	
WHAT	X	X	
CONSEQUENCE	X	X	
COMPARISON			X
COMMENT	X	X	
DECLARATION	X	X	X
GOAL			X
HISTORY	X	X	X
PREDICTION			X
COUNTERMEASURES			
SITUATION			
HOW			

Tabela 5: Abordagens utilizadas para cada *macroaspecto*

4.1. WHAT

Segundo Rassi et al. (2013), o *macroaspecto* WHAT é definido como *"um fato/evento descrito no texto".* Na Figura 6, descreve-se o fato do avião ter sofrido uma queda causando a morte de várias pessoas.

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

Figura 6: Sentença anotada com o *macroaspecto* WHAT

Na Tabela 6, mostra-se a distribuição do aspecto WHAT no córpus. Observa-se que o aspecto WHAT está bem distribuído entre todas as categorias.

Categoria	Frequência
Cotidiano	37
Esportes	24
Mundo	43
Política	62
Total	**166**

Tabela 6: Distribuição do aspecto WHAT por categoria

Na Tabela 7, apresentam-se os resultados do classificador usando os atributos definidos por Teufel (1999). Observa-se que o classificador para a classe "sim" teve melhores resultados do que para classe "não". Para a classe "sim", a cobertura (0.660) foi melhor que a precisão (0.550). Já na Tabela 8, apresentam-se os resultados do melhor classificador usando atributos léxico-semânticos (ver Apêndice C), denominado "*(2, 2) bag_of_words*", criado com base em todos os bigramas "(2,2)" de todas as palavras do córpus. Observa-se que a classe "sim" foi melhor do que a classe "não" com uma alta cobertura (0.800) e uma precisão relativamente baixa (0.519). Cabe ressaltar que os resultados são bons por causa da grande quantidade de sentenças anotadas com WHAT.

WHAT	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	33	17	Sim	0.660	0.550	0.600	0.546
Falso	27	20	Não	0.426	0.541	0.476	

Tabela 7: Resultados do *macroaspecto* WHAT usando atributos de Teufel (1999)

WHAT	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	40	10	Sim	0.800	0.519	0.630	0.515
Falso	37	10	Não	0.213	0.500	0.299	

Tabela 8: Resultados do *macroaspecto* WHAT usando atributos léxico-semânticos

4.2. CONSEQUENCE

Segundo Rassi et al. (2013), o *macroaspecto* CONSEQUENCE é definido como *"um fato/evento causado por outro fato/evento".* Na Figura 7, descreve-se o acontecimento de uma inundação causada pela passagem de um furação.

O furacão Dean passou pela costa sul da Jamaica, inundando a capital e espalhando árvores e telhados.

Figura 7: Sentença anotada com o *macroaspecto* CONSEQUENCE

Na Tabela 9, mostra-se a distribuição do aspecto CONSEQUENCE no córpus. Observa-se a pouca quantidade de instâncias anotadas para todas as categorias.

Categoria	Frequência
Cotidiano	22
Esportes	10
Mundo	12
Política	3
Total	**47**

Tabela 9: Distribuição do aspecto CONSEQUENCE por categoria

Na Tabela 10, apresentam-se os resultados do classificador usando os atributos definidos por Teufel (1999). Observa-se que o classificador para a classe "não" teve melhores resultados do que para classe "sim". Para a classe "sim", tanto a cobertura quanto a precisão são nulos. Os resultados claramente mostram que não é possível identificar CONSEQUENCE usando os atributos de Teufel (1999). Por outro lado, na Tabela 11, mostra-se o melhor classificador usando atributos léxico-semânticos (ver Apêndice C): "*(1, 1) lemmas*", criado com base em

todos os unigramas "(1, 1)" de todas os lemas das palavras do córpus. Para a classe "sim", o classificador obteve uma cobertura quase nula (0.071) e uma precisão perfeita (1.000). Embora o classificador usando atributos léxico-semânticos seja o melhor, não deve ser considerado como um classificador apto para identificar CONSEQUENCE, devido aos resultados quase nulos. Em conclusão, não foi possível identificar o aspecto CONSEQUENCE.

CONSEQUENCE	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	0	14	Sim	0.000	0.000	0.000	0.856
Falso	0	83	Não	1.000	0.856	0.922	

Tabela 10: Resultados do *macroaspecto* CONSEQUENCE usando atributos Teufel (1999)

CONSEQUENCE	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	1	13	Sim	0.071	1.000	0.133	0.866
Falso	0	83	Não	1.000	0.865	0.927	

Tabela 11: Resultados do *macroaspecto* CONSEQUENCE usando atributos léxico-semânticos

4.3. COMPARISON

Segundo Rassi et al. (2013), o *macroaspecto* COMPARISON é definido como *"dados ou estatísticas diferentes comparando duas ou mais entidades".* Na Figura 8, descreve-se a comparação entre autuações em um período atual e em um período do ano anterior.

Foram autuados 208.471 contribuintes, um crescimento de 104,47% em relação ao mesmo período do ano passado.

Figura 8: Sentença anotada com o *macroaspecto* COMPARISON

Na Tabela 12, mostra-se a distribuição do aspecto COMPARISON no córpus. Evidentemente, observa-se a pouca quantidade de instâncias anotadas entre todas as categorias, sendo que *Esportes* e *Mundo* não possuem nenhuma instância anotada.

Categoria	Frequência
Cotidiano	1
Esportes	0
Mundo	0
Política	5
Total	**6**

Tabela 12: Distribuição do aspecto COMPARISON por categoria

Na Tabela 13, apresentam-se os resultados usando regras manuais. Observa-se que a classe "não" teve melhores resultados do que a classe "sim". No entanto, a classe "sim" obteve uma medida F1 relativamente alta (0.667). A acurácia também foi bastante alta (0.991). Os resultados mostram que é possível identificar COMPARISON usando regras manuais (ver Figura 14). Cabe ressaltar que os resultados são bons por causa da pouca quantidade de sentenças anotadas (o que indica um possível *overfitting*).

COMPARISON	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	3	3	Sim	0.500	1.000	0.667	0.991
Falso	0	316	Não	1.000	0.991	0.995	

Tabela 13: Resultados do *macroaspecto* COMPARISON usando regras manuais

4.4. COMMENT

Segundo Rassi et al. (2013), o *macroaspecto* COMMENT é definido como *"um comentário do autor sobre um fato/evento".* Na Figura 9, o autor do texto fez um comentário sobre o esporte brasileiro.

Neste domingo, o esporte brasileiro alegrou a torcida verde-amarelo.

Figura 9: Sentença anotada com o *macroaspecto* COMMENT

Na Tabela 14, mostra-se a distribuição do aspecto COMMENT no córpus. Observa-se a pouca quantidade de instâncias anotadas por categoria, sendo que *Mundo* e *Política* não possuem nenhuma instância anotada.

Categoria	Frequência
Cotidiano	4
Esportes	20
Mundo	0
Política	0
Total	**24**

Tabela 14: Distribuição do aspecto COMMENT por categoria

Na Tabela 15, apresentam-se os resultados do classificador usando os atributos de Teufel (1999). Observa-se que o classificador para a classe "não" teve melhores resultados do que para classe "sim". Para a classe "sim", tanto a cobertura quanto a precisão são nulos. Os resultados afirmam que não é possível identificar COMMENT usando os atributos de Teufel (1999). Já na Tabela 16, apresentam-se os resultados do melhor classificador usando atributos léxico-semânticos: "*(2, 2) semantic*", criado com base em todos os bigramas "(2, 2)" das etiquetas semânticas de todas as palavras do córpus. Para a classe "sim", o classificador obteve uma cobertura baixa (0.143) e uma precisão perfeita (1.000). Mesmo que o classificador com base nos atributos léxico-semânticos seja o melhor (0.025 de F1), não é um classificador competente para identificar COMMENT. Em conclusão, não é possível identificar COMMENT. Cabe ressaltar que os resultados são baixos devido a pouca quantidade de sentenças anotadas com COMMENT.

COMMENT	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	0	7	Sim	0.000	0.000	0.000	0.928
Falso	0	90	Não	1.000	0.928	0.963	

Tabela 15: Resultados do *macroaspecto* COMMENT usando atributos de Teufel (1999)

COMMENT	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	1	6	Sim	0.143	1.000	0.025	0.938
Falso	0	90	Não	1.000	0.938	0.968	

Tabela 16: Resultados do *macroaspecto* COMMENT usando atributos léxico-semânticos

4.5. DECLARATION

Segundo Rassi et al. (2013), o *macroaspecto* DECLARATION é definido como *"um discurso ou fala de alguém ou de uma fonte por citação direta ou indireta".* Na Figura 10, mostra-se a declaração feita por "um informante da delegacia".

> Segundo um informante da delegacia, os dois teriam vendido o acessório de luxo avaliado em cerca de R$10 mil.

Figura 10: Sentença anotada com o *macroaspecto* DECLARATION

Na Tabela 17, mostra-se a distribuição do aspecto DECLARATION no córpus. Observa-se que *Esportes* é a categoria com a menor quantidade de instâncias anotadas.

Categoria	Frequência
Cotidiano	24
Esportes	2
Mundo	14
Política	18
Total	**58**

Tabela 17: Distribuição do aspecto DECLARATION por categoria

Na Tabela 18, apresentam-se os resultados do classificador usando os atributos definidos por Teufel (1999). Nota-se que o classificador para a classe "não" teve melhores resultados do que para a classe "sim". Para a classe "sim", tanto a cobertura quanto a precisão são nulas. Os resultados afirmam que não é possível identificar DECLARATION usando os atributos de Teufel (1999). Na Tabela 19, mostram-se os resultados do melhor classificador usando atributos léxico-semânticos: "*(1, 1) lemmas+POS*", criado com base em todos os unigramas "(1, 1)" do lema junto com a classe gramatical de todas as palavras do córpus. Para a classe "sim", o classificador obteve uma cobertura média (0.529) e uma precisão bastante alta (0.900). Assim, o classificador usando atributos léxico-semânticos obteve os melhores resultados (0.667 de F1).

DECLARATION	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	0	17	Sim	0.000	0.000	0.000	0.825
Falso	0	80	Não	1.000	0.825	0.904	

Tabela 18: Resultados do *macroaspecto* DECLARATION usando atributos de Teufel (1999)

DECLARATION	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	9	8	Sim	0.529	0.900	0.667	0.907
Falso	1	79	Não	0.988	0.908	0.946	

Tabela 19: Resultados do *macroaspecto* DECLARATION usando atributos léxico-semânticos

Na Tabela 20, apresentam-se os resultados usando regras manuais. Observa-se que a classe "não" teve melhores resultados do que a classe "sim" por uma diferença mínima. Tanto a cobertura (0.879) quanto a precisão (0.944) para a classe "sim" foram altas, obtendo-se, por consequência, uma medida F1 bastante alta (0.911). Cabe ressaltar que a acurácia também foi

bastante alta (0.969). Os resultados claramente mostram que é possível identificar DECLARATION usando regras manuais (ver Figura 15).

DECLARATION	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	51	7	Sim	0.879	0.944	0.911	0.969
Falso	3	261	Não	0.989	0.974	0.981	

Tabela 20: Resultados do *macroaspecto* DECLARATION usando regras manuais

4.6. GOAL

Segundo Rassi et al. (2013), o *macroaspecto* GOAL é definido como *"a finalidade/razão para um fato/evento que irá acontecer".* Na Figura 11, mostra-se claramente qual é o objetivo das "buscas".

> O objetivo das buscas é garantir a apreensão dos registros de ocorrências que contêm informações sobre as falhas no controle de tráfego aéreo.

Figura 11: Sentença anotada com o *macroaspecto* GOAL

Na Tabela 21, mostra-se a distribuição do aspecto GOAL no córpus. Claramente pode se observar a mínima quantidade de instâncias anotadas para cada categoria.

Categoria	Frequência
Cotidiano	1
Esportes	1
Mundo	4
Política	4
Total	**10**

Tabela 21: Distribuição do aspecto GOAL por categoria

Na Tabela 22, apresentam-se os resultados usando regras manuais. Observa-se que a classe "não" teve melhores resultados do que a classe "sim". Para a classe "sim", a cobertura foi baixa (0.400), enquanto a precisão foi alta (0.800). Ressalta-se, também, o bom desempenho em termos de acurácia (0.978). Os resultados mostram que é possível identificar GOAL usando regras manuais (ver Figura 11). Também é preciso dizer que foi fácil de se identificar regras por causa da pouca quantidade de instâncias anotadas, e isso pode gerar *overfitting*.

GOAL	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	4	6	Sim	0.400	0.800	0.533	0.978
Falso	1	311	Não	0.997	0.981	0.989	

Tabela 22: Resultados do *macroaspecto* GOAL usando regras manuais

4.7. HISTORY

Segundo Rassi et al. (2013), o *macroaspecto* HISTORY é definido como *"informação de contexto sobre uma história/um passado relacionado ao fato/evento".* Na Figura 12, menciona-se um fato que aconteceu no passado (ano 2002).

Este foi o maior acidente ferroviário egípcio desde 2002, após o incêndio de um trem que deixou 376 mortos.

Figura 12: Sentença anotada com o *macroaspecto* HISTORY

Na Tabela 23, mostra-se a distribuição do aspecto HISTORY no córpus. Observa-se uma baixa quantidade de instâncias anotadas por categoria, sendo *Esportes* e *Política* as categorias com menor quantidade de instâncias anotadas.

Categoria	Frequência
Cotidiano	10
Esportes	4
Mundo	13
Política	2
Total	**29**

Tabela 23: Distribuição do aspecto HISTORY por categoria

Na Tabela 24, apresentam-se os resultados do classificador usando os atributos definidos por Teufel (1999). Nota-se que o classificador para a classe "não" teve melhores resultados do que para a classe "sim". Para a classe "sim", tanto a cobertura quanto a precisão são nulas, portanto, a medida F1 também é nula (0.000). Os resultados mostram claramente que não é possível identificar HISTORY usando os atributos de Teufel (1999). Já na Tabela 28, mostram-se os resultados do classificador usando atributos léxico-semânticos: "*(2, 3) semantic*", criado com base em todos os bigramas e trigramas "(2, 3)" das etiquetas semânticas de todas as palavras do córpus. O classificador obteve uma cobertura bastante baixa (0.111) e uma precisão média (0.500). Embora o classificador baseado em atributos léxico-semânticos tenha obtido os melhores resultados, não é apto para identificar HISTORY, por causa do baixo desempenho. É importante dizer que os resultados são bastante baixos por causa da pouca quantidade de sentenças anotadas com HISTORY.

HISTORY	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	0	17	Sim	0.000	0.000	0.000	0.907
Falso	0	80	Não	1.000	0.907	0.951	

Tabela 24: Resultados do *macroaspecto* HISTORY usando atributos de Teufel (1999)

HISTORY	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	1	8	Sim	0.111	0500	0.182	0.907
Falso	1	87	Não	0.989	0.916	0.951	

Tabela 25: Resultados do *macroaspecto* HISTORY usando atributos léxico-semânticos

Na Tabela 26, apresentam-se os resultados usando regras manuais. Observa-se que a classe "não" teve melhores resultados do que a classe "sim" por uma grande diferença. Para a classe "sim", a cobertura foi relativamente baixa (0.414), enquanto a precisão foi alta (0.750). Cabe ressaltar que a acurácia também foi bastante alta (0.935). Os resultados mostram que é possível identificar HISTORY usando regras manuais (ver Figura 17).

HISTORY	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	12	17	Sim	0.414	0.750	0.533	0.935
Falso	4	289	Não	0.986	0.944	0.965	

Tabela 26: Resultados do *macroaspecto* HISTORY usando regras manuais

4.8. PREDICTION

Segundo Rassi et al. (2013), o *macroaspecto* PREDICTION é definido como *"a informação sobre a factibilidade de fatos/eventos futuros (podendo, inclusive, ser um evento com ocorrência certa)".* Na Figura 13, narra-se a previsão da possível vitória de presidente Lula.

> Com 2 pontos percentuais para mais e para menos, os resultados assegurariam vitória de Lula no primeiro turno.

Figura 13: Sentença anotada com o *macroaspecto* PREDICTION

A abordagem utilizando regras foi avaliada com um total de 322 instâncias (ou sentenças) do córpus. Na Tabela 27, apresenta-se a distribuição do aspecto PREDICTION no córpus. Claramente pode se observar a pouca quantidade de instâncias anotadas para cada categoria.

Categoria	Frequência
Cotidiano	1
Esportes	5
Mundo	6
Política	5
Total	**17**

Tabela 27: Distribuição do aspecto PREDICTION por categoria

Na Tabela 28, apresentam-se os resultados usando regras manuais. Observa-se que a classe "não" teve melhores resultados do que a classe "sim". Para a classe "sim", a cobertura foi alta (0.765), enquanto a precisão foi baixa (0.333). Cabe ressaltar que existe uma grande quantidade de sentenças que deveriam ter sido anotadas manualmente, assim, muitos "falsos positivos" surgiram, ocasionando uma baixa precisão. Ressalta-se, também, o bom desempenho da acurácia (0.907). Os resultados mostram que pode ser factível utilizar regras para identificar PREDICTION (ver Figura 16). Também é preciso dizer que foi fácil de se criar as regras por causa da pouca quantidade de instâncias anotadas, podendo, novamente, gerar *overfitting*.

PREDICTION	Verdadeiro (P)	Falso (P)	Classe	Cobertura	Precisão	F1	Acurácia
Verdadeiro	13	4	Sim	0.765	0.333	0.464	0.907
Falso	26	279	Não	0.915	0.986	0.949	

Tabela 28: Resultados do *macroaspecto* PREDICTION usando regras manuais

5. Conclusões

Nesse relatório, foram apresentados o processo e os resultados da "identificação automática de *macroaspectos*". No total, foram avaliados dois tipos de abordagens: usando AM e usando

regras manuais. A abordagem usando AM visa criar classificadores binários com base nos atributos definidos por Teufel (1999) e atributos léxico-semânticos. Já a abordagem usando regras está baseada nos padrões linguísticos identificados sobre todas sentenças anotadas no córpus. As duas abordagens foram avaliadas sobre o córpus CSTNews (Cardoso et al., 2011).

Os classificadores binários da abordagem AM foram avaliados com base na técnica de estratificação de dados, em que um conjunto de dados é dividido 10 vezes em subconjuntos de *treino* e *teste*, proporcionando uma melhor distribuição das classes. Assim, para cada iteração, os classificadores foram testados com apenas 30% das instâncias (ou sentenças) anotadas do córpus CSTNews, ou seja, um total de 97 sobre 322 sentenças. O resultado final é a média dos valores obtidos em cada iteração. O melhor resultado foi obtido pelo classificador do *macroaspecto* WHAT utilizando atributos léxico-semânticos, por ter um maior número de instâncias anotadas. Também é importante dizer que os atributos definidos por Teufel (1999) são mais apropriados para textos científicos do que para textos jornalísticos. Pode-se acrescentar que o baixo desempenho dos classificadores se deve a pouca quantidade de instâncias anotadas de treino e teste. Acredita-se que a existência de mais instâncias no córpus possa melhorar o desempenho dos classificadores.

Diferentemente da abordagem usando AM, a abordagem utilizando regras foi testada com o córpus anotado completo, ou seja, com um total de 322 sentenças. Os resultados obtidos para alguns *macroaspectos* são razoáveis, provando que é possível identificar macroaspectos usando regras manuais. Um dos grandes fatores pelo qual os resultados não foram maiores é a anotação de aspectos do córpus CSTNews (Rassi et al., 2013). Pode-se perceber, em várias ocasiões, que as regras identificaram automaticamente sentenças que não foram anotadas manualmente (mas que deveriam ter sido anotadas), como aconteceu com PREDICTION, afetando o desempenho das regras (“falsos positivos”).

Para finalizar, os resultados demostraram que é possível identificar certos *macroaspectos* usando técnicas de AM e regras manuais. O aspecto WHAT pode ser identificado usando-se um classificador com base em atributos léxico-semânticos. Já os aspectos COMPARISON, DECLRATION, GOAL, HISTORY e PREDICTION podem ser identificados utilizando-se regras manuais. Por outro lado, tanto o classificador de AM quanto as regras manuais não conseguiram identificar CONSEQUENCE ou COMMENT. Já os aspectos COUNTERMEASURES, SITUATION e HOW não foram avaliados, por terem muito poucas instâncias anotadas no córpus.

Trabalhos futuros incluem a identificação de *microaspectos* e *macroaspectos* para 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

Barrera, Araly, Rakesh Verma, and Ryan Vincent. 2011. “SemQuest: University of houston's semantics'based question answering system.” Paper presented at the 4th Text Analysis Conference, 1-8. Gaithersburg, Maryland, USA. National Institute of Standards and Technology.

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

Bokan, Alessandro, and Thiago A. S. Pardo. 2015. “Identificação Automática de Microaspectos em Textos Jornalísticos.” Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, NILC-TR-15-01, 406:1-43. São Carlos, São Paulo, Brazil, April.

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

Cardoso, Paula, Erick Maziero, Maria Jorge, Eloise Seno, Ariani Di Felippo, Lucia Rino Maria das Graças Nunes, and Thiago 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.

Dayrell, Carmen, Arnaldo Candido, Gabriel Lima, Danilo Machado Jr., Ann Copestake, Valéria Feltrim, Stella Tagnin, and Sandra Aluísio. 2012. “Rhetorical Move Detection in English Abstracts: Multi-label Sentence Classifiers and their Annotated Corpora.” Paper presented at the 8th International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey.

Feltrim, Valeira, Simone Teufel, Maria Nunes, and Sandra Aluísio. 2006. “Argumentative Zoning Applied to Critiquing Novices Scientific Abstracts.” In *Computing Attitude and Affect in Text: Theory and Applications*, volume 20, pp. 233-246. Springer: Information Retrieval Series.

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.

Genoves Jr., Luiz, Valeria Feltrim, Carmen Dayrell, and Sandra Aluísio. 2007. “Automatically Detecting Schematic Structure Components of English Abstracts: Building High Accuracy Classifier for the Task.” Paper presented at the International Workshop on Natural Language Processing for Educational Resources, 23-29. Borovets, Bulgaria.

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.

Makino, Takuya, Hiroya Takamura, and Manabu Okumura. 2012. “Balanced Coverage of Aspects for Text Summarization.” Paper presented at the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, 1742-1746. New York, 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.

Newman, David, Seth Hettich, Cathy Blake, and Christopher Merz. 1998. *UCI Repository of machine learning databases*. University of California, Dept. of Information and Computer Sciences.

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.

Platt, John. 1998. “Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines”. Technical Report of Microsoft Research, MSR-TR-98-14, 1-21. April 21.

Quinlan, Ross. 1993. *C4.5 Programs for Machine Learning*. San Francisco, USA: Morgan Kaufmann Publishers.

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, Brasil, October.

Read, Jesse, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2009. “Classifier Chains for Multi-label Classification.” Paper presented at the Machine Learning and Knowledge Discovery in Databases, Proceedings, 5782:254-269. Bled, Slovenia. September 7-11.

Salton, Gerard, and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. Tokyo: McGraw-Hill.

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. 1999. “Argumentative Zoning: Information Extraction from Scientific Text.” PhD diss., University of Edinburgh.

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.

Tsoumakas, Grigorios, and Ioannis Vlahavas. 2007. “Random k-Labelsets: An Ensemble Method for Multilabel Classification.” Paper presented at the 18th European Conference on Machine Learning (ECML 2007), 4701:406--417. Warsaw, Poland, September 17-21.

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.

Zhou, Liang, Miruna Ticrea, and Eduard Hovy. 2005. “Multi-document Biography Summarization.” Paper presented at the Conference on Empirical Methods in Natural Language Processing, 1-8.

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

Nas tabelas deste apêndice, são listados os aspectos (*microaspectos* e *macroaspectos*) definidos para as 4 categorias do córpus CSTNews: *Cotidiano, Esportes, Mundo e Política*.

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

Tabela 29: 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 30: Aspectos identificados na categoria *Esportes*

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

Tabela 31: 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 32: Aspectos identificados na categoria *Política*

Apêndice B – Regras criadas para identificação de macroaspectos

Nas tabelas deste apêndice, são listadas todas as regras manuais definidas para os *macroaspectos* COMPARISON, DECLARATION, PREDITION, HISTORY e GOAL.

Regra 1: Se a sentença contiver a PREPOSIÇÃO "em", seguida de ("relação"|"comparação"), seguida do ARTIGO "a", então a sentença será anotada como COMPARISON.

"Foram autuados 208.471 contribuintes, um crescimento de 104,47% em relação ao ano passado."
em_PREPOSIÇÃO + relação + a_(ARTIGO)"

Regra 2: Se a sentença tiver o VERBO "comparar", então a sentença será anotada como COMPARISON.

"... as intenções de voto para Lula caíram quando se compara com os candidatos Geraldo e Heloisa."
compara_(VERBO) = comparar_(VERBO)

Figura 14: Regras do *macroaspecto* COMPARISON

verbos_ilocutórios = [dizer, afirmar, anunciar, informar, destacar, expressar, referir, opinar, classificar, admitir, comentar, divulgar]

PESSOA/ORGANIZAÇÃO = [H, Hprof, hum, admin, org, ints, media, party, suborg]

⊂ = "está contido em"

Regra 1: Se a sentença tiver um VERBO contido nos "verbos_ilocutórios", então a sentença será anotada como DECLARATION.

"Marcelinho disse que logo a torcida vai se acostumar e apoiar a mudança de levantador."
disse_(VERBO) = dizer_(VERBO) ⊂ verbos_ilocutórios

"Neste mesmo dia, o exército israelense afirmou ter matado 30 milicianos do Hezbollah."
afirmou_(VERBO) = afirmar_(VERBO) ⊂ verbos_ilocutórios

Regra 2: Se a sentença tiver a PREPOSIÇÃO "segundo", seguida por um ARTIGO, então a sentença será anotada como DECLARATION.

"Segundo o secretário-adjunto da Receita, o Leão está mais atento, não guloso."
segundo_(PREPOSIÇÃO) + o_(ARTIGO)

Regra 3: Se sentença tiver a PREPOSIÇÃO "segundo", seguida por um substantivo associado a uma etiqueta semântica do tipo PESSOA/ORGANZAÇÃO[9], então a sentença será anotada como DECLARATION.

"Segundo Lula, o mundo precisa de uma nova matriz energética, e o etanol pode..."
segundo_(PREPOSIÇÃO) + (Lula_(H) ⊂ PESSOA/ORGANIZAÇÃO)

Regra 3: Se a sentença tiver a PREPOSIÇÃO "de", seguida do VERBO "acordo", seguida da PREPOSIÇÃO "com", então a sentença será anotada como DECLARATION.

"De acordo com a Infraero, será possível realizar a obra em três etapas."
de_PREPOSIÇÃO + acordo_(VERB) + com_(PREPOSIÇÃO)

Figura 15: Regras do *macroaspecto* DECLARATION

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

Regra 1: Se a sentença tiver um VERBO no futuro, então a sentença será anotada como PREDICTION.

"A seleção brasileira ainda enfrentará portugueses e finlandeses na fase de classificação."
*enfrentará*_(VERBO_no_futuro)

Regra 2: Se a sentença tiver o token "previsão", então a sentença será anotada como PREDICTION.

"A previsão de chuva na área aumenta os temores de mais devastação."

Figura 16: Regras do *macroaspecto* PREDICTION

Regra 1: Se a sentença tiver o ADVÉRBIO "já", seguido de um VERBO no tempo pretérito perfeito (PS[10]), pretérito mais que perfeito (PS/MQP), pretérito imperfeito (IMPF) ou condicional (COND); então a sentença será anotada como HISTORY.

"As ações são atribuídas à facção criminosa PCC, que já comandou outros ataques em duas ocasiões."
já_(ADVÉRBIO) + comandou_(PS)

"ACM já tinha sofrido infarto em 1989 e já tinha recebido três pontes de safena."
já_(ADVÉRBIO) + tinha_(IMPF)

Regra 2: Se a sentença tiver o *token* "desde", então a sentença será anotada como HISTORY.

"O grupo criminoso desviou desde 2004 cerca de R$ 70 milhões dos cofres públicos."

"Ele está envolvido com o tráfico desde 1986 e criou sua própria rede distribuidora de drogas..."

Regra 3: Se a sentença tiver a PREPOSIÇÃO+ARTIGO "da", seguido do *token* "história", então a sentença será anotada como HISTORY.

"Um atirador matou ... no pior ataque a tiros contra um campus universitário da história *dos EEUU."*
da_(PREPOSIÇÃO+ARTIGO) + "história"

Figura 17: Regras do *macroaspecto* HISTORY

Regra 1: Se a sentença tiver o *token* lematizado "objetivo", então a sentença será anotada como GOAL.

"O governo israelense objetiva uma zona de segurança cedida a uma força multinacional apoiando..."
objetiva = objetivo_(lema)

"O objetivo das buscas é garantir a apreensão dos registros de ocorrências que contêm informações..."
objetivo = objetivo_(lema)

Figura 18: Regras do *macroaspecto* GOAL

[10] O sistema PALAVRAS fornece etiquetas morfossintáticas para cada *token*/palavra da sentença. Todas as etiquetas encontram-se no link http://beta.visl.sdu.dk/visl/pt/info/portsymbol.html#inflectiontags

Apêndice C – Resultado dos classificadores usando AM

Nas Tabela 33, apresentam-se os resultados da classe “sim” do classificador usando atributos léxico-semânticos para os *macroaspectos* WHAT, CONSEQUENCE, COMMENT, DECLARATION e HISTORY.

Macroaspecto	Classificador	Cobertura	Precisão	F1	Acurácia
WHAT	(1, 1) bag_of_words	0.48+/-0.11	0.632+/-0.1	0.545+/-0.1	0.583+/-0.08
	(2, 2) bag_of_words	**0.8+/-0.69**	**0.519+/-0.18**	**0.63+/-0.39**	**0.515+/-0.03**
	(2, 3) bag_of_words	0.54+/-0.83	0.529+/-0.35	0.535+/-0.45	0.52+/-0.05
	(1, 1) lemmas	0.48+/-0.13	0.6+/-0.06	0.533+/-0.08	0.567+/-0.04
	(2, 2) lemmas	0.72+/-0.75	0.529+/-0.24	0.61+/-0.41	0.526+/-0.05
	(2, 3) lemmas	0.4+/-0.73	0.571+/-0.32	0.471+/-0.39	0.536+/-0.06
	(1, 1) POS	0.46+/-0.09	0.59+/-0.12	0.517+/-0.07	0.561+/-0.08
	(2, 2) POS	0.56+/-0.09	0.609+/-0.1	0.583+/-0.09	0.588+/-0.09
	(2, 3) POS	0.56+/-0.12	0.596+/-0.09	0.577+/-0.09	0.577+/-0.08
	(1, 1) semantic	0.54+/-0.11	0.574+/-0.08	0.557+/-0.07	0.557+/-0.08
	(2, 2) semantic	0.5+/-0.13	0.641+/-0.13	0.562+/-0.11	0.602+/-0.1
	(2, 3) semantic	0.52+/-0.16	0.703+/-0.1	0.598+/-0.12	0.639+/-0.08
	(1, 1) lemmas+POS	0.44+/-0.13	0.579+/-0.06	0.5+/-0.09	0.542+/-0.05
	(2, 2) lemmas+POS	0.62+/-0.76	0.525+/-0.24	0.569+/-0.4	0.515+/-0.07
	(2, 3) lemmas+POS	0.4+/-0.7	0.571+/-0.3	0.471+/-0.38	0.536+/-0.08
	(1, 1) POS+semantic	0.54+/-0.12	0.6+/-0.12	0.568+/-0.09	0.573+/-0.1
	(2, 2) POS+semantic	0.52+/-0.1	0.591+/-0.11	0.553+/-0.09	0.567+/-0.1
	(2, 3) POS+semantic	0.56+/-0.1	0.596+/-0.11	0.577+/-0.08	0.577+/-0.09
CONSEQUENCE	(1, 1) bag_of_words	0.071+/-0.09	1.0+/-0.87	0.133+/-0.15	0.866+/-0.02
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(1, 1) lemmas	**0.071+/-0.09**	**1.0+/-0.81**	**0.133+/-0.16**	**0.866+/-0.02**
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(2, 2) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.01
	(2, 3) POS	0.071+/-0.15	0.143+/-0.28	0.095+/-0.19	0.804+/-0.05
	(1, 1) semantic	0.071+/-0.17	0.333+/-0.53	0.118+/-0.25	0.845+/-0.02
	(2, 2) semantic	0.071+/-0.09	0.5+/-0.8	0.125+/-0.16	0.856+/-0.02
	(2, 3) semantic	0.071+/-0.15	0.333+/-0.75	0.118+/-0.24	0.845+/-0.04
	(1, 1) lemmas+POS	0.071+/-0.11	1.0+/-0.77	0.133+/-0.18	0.866+/-0.01
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.856+/-0.0
	(1, 1) POS+semantic	0.0+/-0.07	0.0+/-0.46	0.0+/-0.12	0.845+/-0.01
	(2, 2) POS+semantic	0.071+/-0.09	0.333+/-0.37	0.118+/-0.15	0.844+/-0.04
	(2, 3) POS+semantic	0.071+/-0.13	0.25+/-0.35	0.111+/-0.18	0.835+/-0.05
COMMENT	(1, 1) bag_of_words	0.143+/-0.25	0.5+/-0.87	0.222+/-0.34	0.928+/-0.02
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0

	(1, 1) lemmas	0.143+/-0.24	0.5+/-0.83	0.222+/-0.35	0.928+/-0.02
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(2, 2) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.01
	(2, 3) POS	0.143+/-0.13	0.25+/-0.3	0.182+/-0.17	0.907+/-0.03
	(1, 1) semantic	0.143+/-0.17	0.5+/-0.72	0.222+/-0.25	0.928+/-0.02
	(2, 2) semantic	**0.143+/-0.18**	**1.0+/-0.84**	**0.25+/-0.28**	**0.938+/-0.01**
	(2, 3) semantic	0.143+/-0.22	1.0+/-0.72	0.25+/-0.32	0.938+/-0.01
	(1, 1) lemmas+POS	0.143+/-0.24	0.5+/-0.78	0.222+/-0.32	0.928+/-0.02
	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.928+/-0.0
	(1, 1) POS+semantic	0.0+/-0.11	0.0+/-0.61	0.0+/-0.18	0.928+/-0.01
	(2, 2) POS+semantic	0.143+/-0.19	0.333+/-0.47	0.2+/-0.27	0.918+/-0.02
	(2, 3) POS+semantic	0.143+/-0.22	0.333+/-0.67	0.2+/-0.3	0.918+/-0.03
DECLARATION	(1, 1) bag_of_words	0.176+/-0.15	0.6+/-0.29	0.273+/-0.2	0.835+/-0.04
	(2, 2) bag_of_words	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(2, 3) bag_of_words	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(1, 1) lemmas	0.412+/-0.24	0.875+/-0.25	0.56+/-0.23	0.887+/-0.05
	(2, 2) lemmas	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(2, 3) lemmas	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(1, 1) POS	0.0+/-0.04	0.0+/-0.6	0.0+/-0.07	0.825+/-0.02
	(2, 2) POS	0.294+/-0.18	0.5+/-0.22	0.37+/-0.19	0.825+/-0.05
	(2, 3) POS	0.353+/-0.19	0.462+/-0.21	0.4+/-0.17	0.814+/-0.06
	(1, 1) semantic	0.353+/-0.22	0.6+/-0.31	0.444+/-0.25	0.845+/-0.05
	(2, 2) semantic	0.235+/-0.24	0.571+/-0.51	0.333+/-0.3	0.835+/-0.05
	(2, 3) semantic	0.176+/-0.24	0.5+/-0.51	0.261+/-0.27	0.825+/-0.05
	(1, 1) lemmas+POS	**0.529+/-0.24**	**0.9+/-0.13**	**0.667+/-0.21**	**0.907+/-0.04**
	(2, 2) lemmas+POS	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(2, 3) lemmas+POS	0.118+/-0.11	1.0+/-0.6	0.211+/-0.19	0.845+/-0.02
	(1, 1) POS+semantic	0.125+/-0.18	0.333+/-0.35	0.182+/-0.22	0.812+/-0.04
	(2, 2) POS+semantic	0.294+/-0.19	0.417+/-0.19	0.345+/-0.17	0.804+/-0.04
	(2, 3) POS+semantic	0.294+/-0.2	0.385+/-0.14	0.333+/-0.15	0.794+/-0.04
HISTORY	(1, 1) bag_of_words	0.0+/-0.07	0.0+/-0.6	0.0+/-0.12	0.907+/-0.01
	(2, 2) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(2, 3) bag_of_words	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(1, 1) lemmas	0.0+/-0.11	0.0+/-0.98	0.0+/-0.2	0.907+/-0.01
	(2, 2) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(2, 3) lemmas	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(1, 1) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(2, 2) POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.01
	(2, 3) POS	0.0+/-0.09	0.0+/-0.61	0.0+/-0.15	0.866+/-0.05
	(1, 1) semantic	0.111+/-0.15	0.5+/-0.69	0.182+/-0.23	0.907+/-0.02
	(2, 2) semantic	0.0+/-0.07	0.0+/-0.6	0.0+/-0.12	0.907+/-0.01
	(2, 3) semantic	**0.111+/-0.11**	**0.5+/-0.91**	**0.182+/-0.19**	**0.907+/-0.02**
	(1, 1) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0

	(2, 2) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(2, 3) lemmas+POS	0.0+/-0.0	0.0+/-0.0	0.0+/-0.0	0.907+/-0.0
	(1, 1) POS+semantic	0.0+/-0.1	0.0+/-0.81	0.0+/-0.18	0.907+/-0.01
	(2, 2) POS+semantic	0.111+/-0.14	0.5+/-0.91	0.182+/-0.23	0.907+/-0.02
	(2, 3) POS+semantic	0.111+/-0.2	0.333+/-0.6	0.167+/-0.26	0.897+/-0.04

Tabela 33: Resultados dos classificadores usando atributos léxico-semânticos