Reconhecimento de Entidades Mencionadas em Notificações de Atos de Concentração do Conselho Administrativo de Defesa Econômica

Renan Fichberg

Orientador: Prof. Dr. Marcelo Finger

Universidade de São Paulo, Instituto de Matemática e Estatística

renan.fichberg@usp.br -- https://linux.ime.usp.br/~fichberg/mac0499/

Introdução

O Conselho Administrativo de Defesa Econômica (CADE) é um orgão independente que reporta ao Ministério da Justiça e possui como missão garantir a livre concorrência de mercado em todo o território Brasileiro e realiza as suas funções legais de acordo com a Lei Nº 12.529/2011 [1]. O CADE dispõem de uma base de dados bastante extensa, com processos judiciais de vários tipos distintos datados do ano de 1980 até os dias atuais. De tais processos, denominados Atos de Concentração, buscamos extrair informações a partir do reconhecimento de entidades mencionadas para futuramente tentarmos descobrir qual dos dois ritos um Ato de Concentração futuro seguirá: sumário ou ordinário.

Processamento de Linguagem Natural

Um dos problemas de processamento de linguagem natural envolvem o entendimento de linguagens naturais por parte das máquinas. Com esta finalidade foi desenvolvido um córpus (um conjunto de textos selecionados), posteriormente anotado manualmente, conforme a Figura 1:

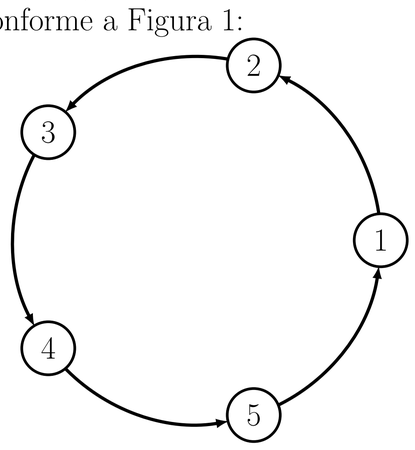

Figura 1: Ciclo de geração de anotações.

1: Leitura e compreensão do documento.
2: Identificação e declaração das EMs.
3: Identificação e declaração dos relacionamentos.
4: Anotar o documento.
5: Aplicação de métricas e validação.

Na construção do córpus, foram considerados o **idioma**, a **estrutura do texto** e a **representatividade** [2] dos dados além do seu **tamanho** final.

Métricas

Foram utilizadas para medir o desempenho do córpus as seguintes métricas tradicionais em recuperação de informações [5]:

A precisão (**P**):

$$\mathbf{P} = \frac{\#\text{itens relevantes recuperados}}{\#\text{itens recuperados}} = \frac{\mathbf{VP}}{\mathbf{VP} + \mathbf{FP}}$$

A cobertura (**C**):

$$\mathbf{C} = \frac{\#\text{itens relevantes recuperados}}{\#\text{itens relevantes}} = \frac{\mathbf{VP}}{\mathbf{VP} + \mathbf{FN}}$$

A medida-F balanceada ($\mathbf{F}_1$):

$$\mathbf{F}_1 = \frac{2\mathbf{PC}}{\mathbf{P} + \mathbf{C}}$$

Onde os códigos **VP**, **FP**, **FN** e **VN** significam Verdeiro Positivo, Falso Positivo, Falso Negativo e Verdeiro Negativo, respectivamente.

Atos de Concentração

Os Atos de Concentração Econômicas (AC) são caracterizados por operações que envolvem duas ou mais empresas independentes, conforme descrito no artigo 90 da Lei 12.529/2011 [2]. As operações mais comuns dos ACs são:

Fusão **Incorporação** **Aquisição** **Joint Venture**

Reconhecimento de Entidades Mencionadas

Uma entidade mencionada (EM) é um objeto do mundo real que possui um nome próprio, como por exemplo uma pessoa ou uma organização. As entidades mencionadas do córpus foram anotadas por meio de uma ferramenta *web* chamada BRAT [3], v.1.3.0, um projeto *open source* (Licença MIT) recente, desenvolvido colaborativamente por pesquisadores de vários grupos distintos com interesse em anotações de texto. Na Figura 1 abaixo, anotações BRAT em uma sentença de um Ato de Concentração:

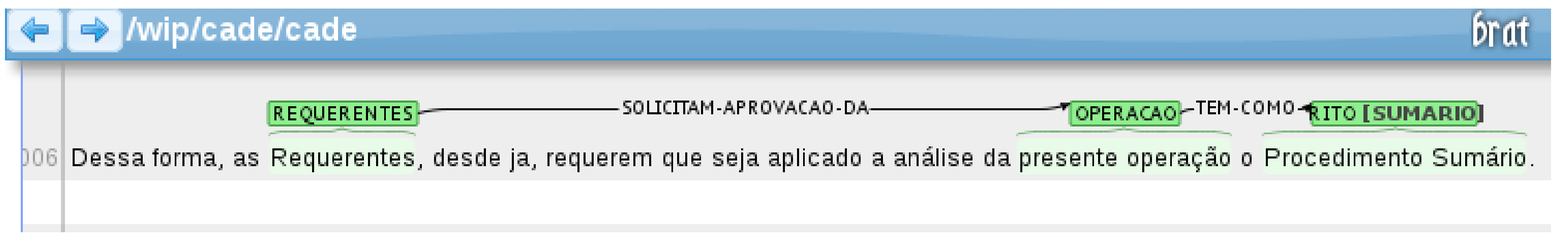

Figura 2: Anotações manuais de entidades mencionadas no BRAT de um dado Ato de Concentração.

Foram usados também os módulos de treinamento e reconhecimento de entidades mencionadas do Apache OpenNLP v.1.6.0 [4]. Na Figura 2, exemplo de saída de reconhecimento de entidades mencionadas de forma automatizada.

```
10. A fim de instruir a presente petição, as Requerentes apresentam os documentos
listados abaixo:
10. A fim de instruir a presente petição, as <START:REQUERENTES> Requerentes <END>
 apresentam os <START:DOCUMENTO> documentos listados <END> abaixo:
```

Figura 3: Reconhecimento de entidades mencionadas pelo Apache OpenNLP para a sentença dada.

Resultados

Resultados da validação cruzada com o método *holdout*. Cada C_i representa uma combinação de grupos de teste e de treinamento compostos por cinco e quarenta e cinco Atos de Concentração, respectivamente. Os rótulos L são apresentados na primeira coluna.

L \ C_i	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
R	327	263	218	324	293	210	270	263	220	243
T	774	527	491	741	653	520	702	623	587	605
A	258	208	168	248	231	162	208	210	179	184
E1 + E3	461 + 55	272 + 47	286 + 37	431 + 62	379 + 43	323 + 35	443 + 51	371 + 42	375 + 33	375 + 46
E2 + E3	14 + 55	8 + 47	13 + 37	14 + 62	19 + 43	13 + 35	11 + 51	11 + 42	8 + 33	13 + 46
P	0.789	0.790	0.770	0.765	0.788	0.771	0.770	0.798	0.813	0.757
C	0.333	0.394	0.342	0.334	0.353	0.311	0.296	0.337	0.305	0.304
$\mathbf{F}_1$	0.468	0.525	0.473	0.465	0.487	0.443	0.427	0.473	0.443	0.433

R: EMs recuperadas
T: EMs existentes na coleção
A: EMs corretamente recuperadas (VP)
E1: EMs perdidas (FN)
E2: EMs classificadas erradas (FP)
E3: EMs imprecisas (FN e FP)
P: Valor precisão
C: Valor cobertura
F_1: Valor medida-F balanceada

Referências

[1] Acesso à Informação: Conheça o CADE. Disponível em: ⟨ http://www.cade.gov.br/acesso-a-informacao/institucional ⟩.

[2] MANNING, Christopher D., SCHUETZE, Hinrich. Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

[3] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. In Proceedings of the Demonstrations Session at EACL 2012.

[4] Documentação Apache OpenNLP. Disponível em: ⟨ https://opennlp.apache.org/documentation/1.6.0/manual/opennlp.html ⟩.

[5] MANNING, Christopher D., RAGHAVAN, Prabhakar., SCHUETZE, Hinrich. An Introduction to Information Retrieval. Online Edition. Cambridge University Press. 2008.

LaTeX TikZposter