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. 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 e mais de 100 tipos diferentes de documentos, desde formulários, notificações de processos e cópias escaneadas de documentos diversos até arquivos de áudio e vídeo.

Os processos judiciais objetos de estudo deste trabalho são os denominados Atos de Concentração. Um dado Ato de Concentração submetido ao Conselho Administrativo de Defesa Econômica pode ser analisado sob dois diferentes ritos: sumário ou ordinário. Em particular, o objetivo do trabalho desenvolvido foi de extrair o máximo de informações potencialmente relevantes para uma futura tentativa de classificação automatizada do tipo de rito que um dado futuro Ato de Concentração poderá seguir. Para isso, foi desenvolvido um córpus composto de cinqüenta Atos de Concentração do ano de 2016 com anotações de entidades mencionadas predefinidas usando conhecimento e técnicas de processamento de linguagem natural e ferramentas de aprendizado de máquina supervisionado.

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. Tais operações podem ser aquisições de controle ou incorporações de uma ou mais empresas por outras ou ainda a celebração de contratos associativos, consórcios ou *joint ventures* entre empresas. A natureza destas operações, aliadas ao faturamento bruto anual ou volume de negócios no Brasil dos agentes econômicos envolvidos, que justificam a existência dos ACs analisados pelo CADE. Quando o faturamento bruto anual de uma das empresas envolvidas na operação atinge o patamar mínimo de R$ 750 milhões e o de uma outra, também envolvida na operação, o patamar de R$ 75 milhões, o AC deve ser notificado ao CADE. Os possíveis tipos de operações que um AC pode ter são:

- **Fusão**: união de duas ou mais empresas distintas para formar um novo agente econômico único.
- **Incorporação**: ato de uma ou mais empresas incorporar total ou parcialmente outras empresas dentro de uma mesma pessoa jurídica, de tal forma que o incorporado desaparece como pessoa jurídica, mas o incorporador mantém a sua identidade jurídica após a operação.
- **Aquisição**: ato de uma empresa adquirir o controle total ou parcial da participação acionária de outra empresa.
- ***Joint venture***: criação de uma nova empresa a partir da associação entre duas ou mais empresas, de tal forma que as empresas que se associaram mantém normalmente suas identidades jurídicas pós operação.

Córpus

Problemas de processamento de linguagem natural envolvem o entendimento de linguagens naturais por parte das máquinas ou mesmo geração de linguagem natural (isto é, a conversão de uma representação entendida por computadores em uma representação em linguagem natural). Este trabalho se encaixa na primeira categoria de tal modo que foi desenvolvido um córpus (um conjunto de textos selecionados), posteriormente anotado manualmente, para gerar um modelo de treinamento capaz de extrair informações de forma automatizada. Foram considerados os seguintes pontos na criação do córpus:

- **Linguagem**: No caso, português.
- **Estrutura**: Forma que as notificações dos Atos de Concentração são estruturadas.
- **Representatividade**: Os Atos selecionados precisam constituir de uma amostra representativa da população de interesse.
- **Tamanho**: Precisa ser suficientemente extenso. O córpus construído tem pouco mais de 50000 palavras.

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, 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:

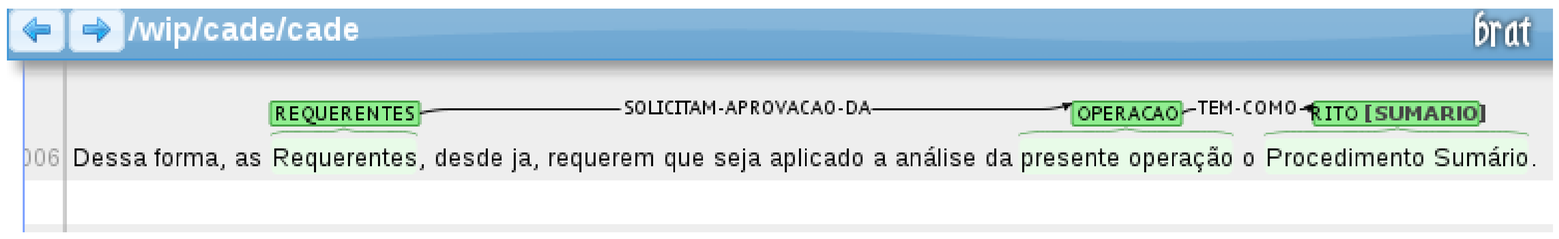

Figura 1: Anotações manuais de entidades mencionadas no BRAT de um dado Ato de Concentração.

Foram usados também os módulos de treinamento e reconhecimento de entidades mencionadas do Apache OpenNLP v.1.6.0. Porém, devido à necessidade de modelos de tokenização e de detecção de sentenças para poder treinar um modelo de reconhecimento de entidades mencionadas, foram usados modelos previamente gerados sobre o córpus Amazônia (possivelmente o córpus em português mais revisado por especialistas e bastante extenso).

O processo de tokenização consiste em quebrar o texto em *tokens* (palavras, números e pontuações, basicamente) enquanto o de detecção de sentenças consiste em quebrar o texto em sentenças, como os nomes sugerem. Assim, a sentença apresentada na Figura 1 acima é tokenizada da seguinte maneira:

"Dessa | forma | , | as | Requerentes | , | desde | já | , | requerem | que | seja | aplicado | a | análise | da | presente | operação | o | Procedimento | Sumário | ."

Métricas

Foram utilizadas para medir o desempenho do córpus as seguintes métricas:

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

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