Espacialização de notícias de cana-de-açúcar utilizando entidades do IBGE

Ercilia Souza Rodrigues[1]
Eduardo Antonio Speranza[2]
Maria Fernanda Moura[2]
Rosa Nathalie Portugal Vargas[3]
Solange de Oliveira Rezende[3]

Os estudos da Divisão Regional do Instituto Brasileiro de Geografia e Estatística (IBGE) tiveram início em 1941, tendo, como principal objetivo, sistematizar divisões regionais que vinham surgindo no país, organizando, assim, uma única divisão regional para a divulgação das estatísticas do país. Atualmente, o Brasil é dividido em cinco regiões segundo a classificação do IBGE (IBGE, 2011), sendo elas: Norte, Nordeste, Centro-Oeste, Sul e Sudeste. Porém, existe outra maneira de classificar o país, captando melhor a situação socioeconômica e as relações entre sociedade e o espaço natural. Trata-se da divisão do país em três grandes complexos regionais, as macrorregiões: Centro Sul, Nordeste e Amazônia.

Outras classificações estabelecidas pelo IBGE que dividem o território são as mesorregiões, que são áreas com características socioeconômicas comuns, utilizadas para fins estatísticos. As Microrregiões são agrupamentos de municípios limítrofes com finalidade de integrar a organização e o planejamento de municípios com similaridades econômicas e sociais.

O foco deste trabalho é gerar a classificação manual de notícias agrícolas, de acordo com essas divisões geográficas, para serem utilizadas no âmbito do projeto “Tecnologias Inovadoras em mineração de textos para a Espacialização de Notícias Agrícolas” (TIENA). O objetivo dessa classifica-

[1] *Pontifícia Universidade Católica de Campinas, ercliasr@cnptia.embrapa.br*
[2] *Embrapa Informática Agropecuária, {speranza, fernanda}@cnptia.embrapa.br*
[3] *Universidade de São Paulo, {nathalie, solange}@icmc.usp.br*

ção é comparar resultados de classificação automática obtida por classificadores desenvolvidos no âmbito do projeto com os resultados manuais, a fim de validar os resultados automaticamente obtidos e encontrar formas de aprimorá-los

Para realizar a classificação de notícias agrícolas, o primeiro passo é formar um Corpus de um domínio fixo de conhecimento; nesse caso, foram utilizadas notícias sobre cana-de-açúcar. Sardinha (2004) descreve *Corpus* como uma abordagem que se ocupa da coleta e exploração de *corpora* ou conjunto de dados linguísticos textuais que foram coletados criteriosamente, com o propósito de servir de ferramenta para uma pesquisa.

Após a montagem do *Corpus* e com um programa, desenvolvido no âmbito do projeto, é possível ler cada notícia do *Corpus* e associá-la a uma classificação, segundo a divisão regional proposta pelo IBGE, em macrorregião, região, mesorregião, microrregião, município, ou usina à qual a notícia pertence. A decisão de qual ou quais classificações devam ser aplicadas a cada noticia é realizada pelo especialista do domínio, no caso um geógrafo com algum conhecimento de produção de cana-de-açúcar, no Brasil. Conferida, subjetivamente, a classificação aplicada, esta é anotada no *Corpus*. Quando todo o *Corpus*, ou uma janela do dele, é, subjetivamente, considerado corretamente marcado, é repassado aos processos de classificação automática.

Após a classificação automática da janela do *Corpus*, os resultados são comparados com os manualmente obtidos. A comparação pode ser completamente subjetiva ou automática. Então, as diferenças consideradas significativas são estudadas caso a caso, a fim de verificar a origem do erro e como tratá-lo.

Os primeiros experimentos foram realizados sobre uma janela de 237 notí-cias. O objetivo foi validar o uso da ferramenta linguística Rembrandt que utiliza um vocabulário controlado para reconhecer automaticamente todas as entidades geo-gráficas em cada notícia. O foco foi apenas nas entidades reconhecidas como de “localização geográfica”. Porém, como exemplificado na Figura 1, Cosan aparece como empresa, mas é uma importante dica de localização geográfica da notícia. Logo, nesses primeiros experimentos, verificamos que o especialista deveria marcar localizações com base em todas as que aparecessem explicitamente na notícia ou que fossem possíveis de localizar via alguma instituição. Além disso, após a identifi-cação automática das entidades, há um processo de desambigua-

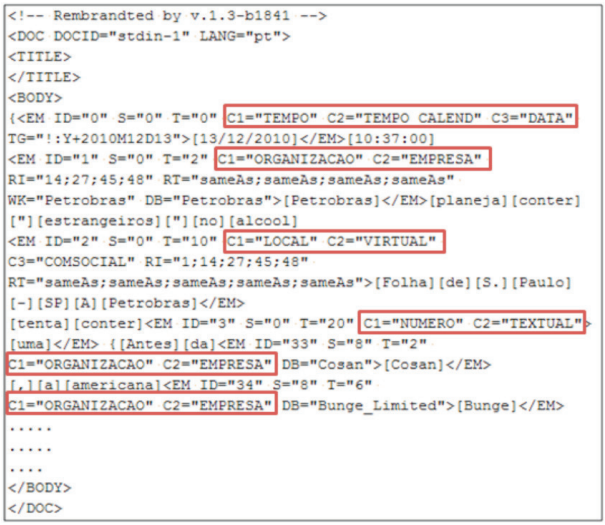

Figura 1. Exemplo de classificação por software, onde são encontradas entidades como tempo, organização (empresa) e local.

ção das classifica-ções, implementado no âmbito do projeto. Por exemplo, para decidir se Belém está aqui ou em algum outro estado cujo nome da cidade também seja Belém, essa de-sambiguação é feita pelo uso das coordenadas geográficas de cada entidade.

Nos trabalhos futuros prevê-se melhorar a classificação manual produzida na janela de 237 notícias, corrigindo erros que foram surgindo com o decorrer da classificação, como a inserção do estado do Mato Grosso na macrorregião Centro Sul. Nesse caso, apenas uma parte do estado se localiza nela, pois a maioria de sua área está na Amazônia. Com isso, espera-se verificar se o Rembrandt identifica as mesmas categorias e se o processo de desambiguação aplicado após essa identificação permite melhor aproximação dos resultados manuais dos automáticos.

Referências

SARDINHA, T. B. **Linguística de Corpus**. Barueri: Manole, 2004. 2004. 410 p.

IBGE. **Instituto Brasileiro de Geografia e Estatística**. 2011. Disponível em: <http://www.ibge.gov.br/home>. Acesso em: 20 out. 2011.