Extração de dados de fontes textuais: uma abordagem para enriquecimento de dados abertos interligados

Karen Torres Teixeira, Maria Luiza Machado Campos, João C. P. da Silva

[1]Programa de Pós-Graduação em Informática
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de janeiro – RJ – Brasil

karentteixeira@gmail.com;mluiza@ppgi.ufrj.br; jcps@dcc.ufrj.br

Abstract. *In the Web of Data, data items are interconnected and associated with descriptive annotations, taking advantage of a representation in the form of triples. In this context, documents and other textual sources can be annotated to be incorporated into this universe as resources or serving as sources for extracting new triples. The purpose of this article is to present an approach for data extraction and triple generation from texts with specific styles, aiming at their association and connection to existing databases. The approach was applied and evaluated in the context of a portal with information on the consumption of pesticides in Brazil.*

Resumo. *Na Web de Dados, itens de dados são interconectados e associados a anotações descritivas na forma de vocabulários, tirando vantagem de uma representação em triplas. Neste contexto, documentos e outras fontes textuais podem ser anotados para serem incorporados a este universo como recursos ou servindo também de base para extração de novas triplas. O objetivo deste artigo é apresentar uma abordagem para extração de dados e geração de triplas a partir de textos com estilos específicos visando o enriquecimento de dados abertos interligados, através de sua associação e ligação a bases existentes. A abordagem foi aplicada e avaliada no contexto de um portal com informações sobre o consumo de agrotóxicos no Brasil.*

1. Introdução

A importância de obter informações e gerar conhecimento é uma das principais motivações para uso de tecnologia computacional no mundo atual. A informação é gerada a partir da disponibilização, organização e exploração de dados, cujo volume cresceu enormemente com o surgimento da *web*, onde uma grande quantidade de dados é publicada na forma de dados abertos.

Nos últimos anos, para permitir que esses dados possam ser explorados de forma conjunta e processados por agentes de software com maior agilidade, surge a assim chamada Web de Dados. Nesta, ao invés de associações entre páginas e documentos, itens de dados são interligados, tirando vantagem de uma representação em triplas, onde os dados são interconectados e associados a anotações descritivas na forma de vocabulários ou ontologias.

Uma consequência importante da interligação de dados seguindo os padrões propostos pela Web Semântica[1] (chamados de dados abertos interligados ou Linked Open

[1]http://www.w3c.br/Padroes/WebSemantica

Data – LOD, em inglês) é a possibilidade de geração de novos conhecimentos a partir da exploração da ligações entre diferentes recursos na *web*.

É neste contexto, que o Observatório de Atenção Permanente ao Uso de Agrotóxicos [2] foi criado, visando disponibilizar informações referentes ao uso de agrotóxicos em um portal de dados abertos. Este portal foi desenvolvido utilizando a plataforma CKAN, que é um gerenciador de conteúdo (dados). O CKAN é muito utilizado por governos, organizações e instituições que coletam muitos dados, por facilitar a realização de tarefas como gerenciamento e publicação dados, busca facetada para navegação e visualização dos dados. Apesar dos dados estarem publicados no portal, eles ainda não se encontram totalmente triplificados (tarefa em andamento), o que não permite sua ligação com outras bases de dados. Para evidenciar um dos benefícios do uso de LOD neste contexto, considere um conjunto de dados que trata de agrotóxicos permitidos e proibidos, e outro que compreende agrotóxicos utilizados em uma determinada produção agrícola. Se esses dados são publicados como LOD, é possível descobrir que um agrotóxico usado intensivamente na produção de tomate é o mesmo agrotóxico descrito no outro dataset e que consta como proibido. A interligação desses itens de dados permitirá não só sua recuperação conjunta, mas também expandirá as possibilidades de descoberta de novos conhecimentos a partir dessas fontes de dados.

Embora exista amplo uso de LOD sobre dados estruturados, como CSV, XML, HTML, o mesmo não vale para dados não estruturados como imagens, vídeos e textos. Isto porque tratar dados não estruturados é uma tarefa árdua. Considerando a enorme quantidade de informações relevantes em formato textual, é muito importante a sua transformação para o formato estruturado ou semiestruturado. A extração de informação de um texto pode gerar um enriquecimento de informação muito maior.

O Processamento de Linguagem Natural (PLN) é uma área da Computação que estuda técnicas e mecanismos que possibilitam ao computador manipular e interpretar os dados não estruturados. Entre as tarefas estudadas em PLN estão o reconhecimento de entidades nomeadas e a extração de relações entre as entidades [Nadkarni et al. 2011]. Essas tarefas auxiliam o computador na identificação de lugares, pessoas, organizações e outras formas de entidades e na relação entre elas. Assim, as informações que antes não possuíam formato estruturado, agora possuem uma representação dos dados que permite sua melhor manipulação e interoperabilidade pelo computador.

Para extrair informações de textos, diversos métodos já foram propostos, podendo ser baseados em (i) regras, (ii) aprendizado supervisionado, (iii) aprendizado semi-supervisionado e (iv) aprendizado não supervisionado [de Abreu et al. 2013].

Esses métodos podem ser utilizados tanto em domínio específico, a exemplo de textos com estilos semelhantes em um determinado domínio, quanto em domínio aberto, onde independem do estilo e domínio do texto e da quantidade de relações existentes. Neste último, não é necessária a análise prévia do texto, sendo utilizadas técnicas referentes à estrutura da linguagem e não referente a um domínio restrito para que ocorra a extração.

É neste contexto de extração de informação, PLN e Web Semântica que este trabalho se insere. O objetivo deste artigo é apresentar uma abordagem para extração de

[2] `http://dados.contraosagrotoxicos.org/pt_PT/`

dados e geração de triplas a partir de textos com estilos específicos, visando o enriquecimento de dados abertos interligados, através de sua associação e ligação a bases existentes. O trabalho foi desenvolvido tendo como alvo os datasets do portal do Observatório e interligando-o, como exemplo, ao Agrovoc [Caracciolo et al. 2013], Bioportal [Noy et al. 2009] e DBpedia [Lehmann et al. 2015].

Este artigo está organizado da seguinte maneira: a seção 2 revisa e classifica trabalhos relacionados a técnicas para extração de informações a partir de textos. A seção 3 descreve a abordagem proposta. Na seção 4 são detalhados os experimentos e os primeiros resultados, finalizando com a seção 5 que apresenta algumas conclusões e os próximos passos.

2. Trabalhos relacionados

Buscando verificar o estado da arte de técnicas e mecanismos para extração de dados a partir de fontes textuais, foi realizado um levantamento de trabalhos nesta área. Esta seção apresenta uma visão geral desse levantamento e uma caracterização dos métodos revisados.

2.1. Método baseado em regras

Esta foi uma das primeiras abordagens propostas para sistemas de extração. Nela é necessário que um humano escreva e desenvolva regras ou expressões regulares para que a informação desejada seja extraída.

Os primeiros trabalhos que utilizaram esta abordagem focaram em domínios específicos. Como exemplo, nos trabalhos de [Grishman et al. 1991] e [Lehnert et al. 1991] as regras foram definidas a partir do domínio e estilo dos textos, sendo seu objetivo extrair informações de textos sobre ações de terrorismo para completar templates com essas informações sobre o tipo do evento, data, localização, vítimas e alvos físicos. Em caso de nova informação, uma nova regra deve ser definida. Esses sistemas não são escaláveis para outros domínios sem precisar de alteração, pois utilizam regras e templates específicos.

A vantagem desta abordagem para domínios específicos é que a semelhança dos estilos de textos facilita a definição das regras. A desvantagem deste método é que, como as regras são pré-definidas, se um novo tipo de relação tiver que ser extraído, então é preciso adicionar uma nova regra ao conjunto existente.

A extração em domínios abertos (Open Information Extraction - OIE) [Etzioni et al. 2008] como a *web* é mais complexa devido à grande quantidade de entidades e relações. Um sistema baseado em OIE deve operar em duas fases. Primeiro, deve aprender um modelo geral de como os relacionamentos e entidades são representados e expressados em uma determinada linguagem. Segundo, deve utilizar esse modelo como base para o extrator de relação independente de domínio, em que a entrada é um corpus de documentos e a saída são triplas extraídas.

Para aprender o modelo geral de como as relações são expressas em uma determinada linguagem, diferentes recursos podem ser utilizados, como por exemplo análise morfológica, sintática e semântica. Podemos citar como exemplos deste tipo de sistema: Graphia [Carvalho et al. 2013], LODifier

[Augenstein et al. 2012], ReVerb [Fader et al. 2011], DepOE [Gamallo et al. 2012] e ClausIE [Del Corro and Gemulla 2013]. Uma característica comum a tais sistemas é que os mecanismos de PLN usados na análise das sentenças têm um impacto forte na qualidade das extrações obtidas.

2.2. Método baseado em aprendizado supervisionado

Esta abordagem surgiu com o objetivo de que o computador pudesse aprender/definir as regras de extração, ao invés de um humano ter que fazê-lo. Esse processo de aprendizado é feito com um conjunto de textos anotados por especialistas de domínio que servem como um conjunto de treinamento para que a máquina possa aprender as regras desejadas. Esta abordagem pode ser aplicada para domínios específicos (como em [Lange et al. 2010] e [Joshi et al. 2013]) ou para domínios abertos (como em [Byrne and Klein 2010] e [de Souza and Claro 2014]).

2.3. Método baseado em aprendizado semi-supervisionado

Esta abordagem tenta contornar o problema dos métodos supervisionados, de necessitar de uma base anotada, utilizando apenas um pequeno conjunto de dados anotados e outro conjunto não-anotado. Exemplos de sistemas que usam essa abordagem são o DIPRE [Brin 1999] e o Snowball [Agichtein and Gravano 2000], ambos aplicados a domínios específicos.

2.4. Método baseado em aprendizado não-supervisionado

Nesta abordagem, o objetivo é extrair informação sem a necessidade de se usar uma base previamente anotada. Isso pode ser feito através de um conjunto de relações genéricas para descobrir/aprender novas relações, atributos e instâncias, ou através da utilização de uma base de conhecimento para o domínio, onde as relações existentes em tais bases possam ser exportadas para outras bases que se refiram ao mesmo domínio da base original.

No sistema Espresso [Pantel and Pennacchiotti 2006], este método foi utilizado. O sistema começa com um conjunto de instâncias e tenta encontrar no texto trechos semelhantes ao conjunto inicial. Neste caso, a extração ocorreu em domínio específico e as relações semânticas extraídas foram referentes a *isa, part-of, succession, reaction, e production*. Outros trabalhos que utilizaram a abordagem não-supervisionada para domínios abertos foram o KnowitAll, [Etzioni et al. 2005], TextRunner [Yates et al. 2007].

Esta abordagem apesar de não necessitar de dados anotados e nem regras escritas por humanos, possui mais chances de extrair relações indesejáveis e incoerentes, pois a base utilizada pode não possuir um conjunto inicial confiável de treinamento.

3. Abordagem proposta

Muitos dos trabalhos apresentados na seção anterior já evidenciaram os benefícios da extração de informações a partir de texto. A escolha do método de extração adequado depende de diversos fatores como, por exemplo, o estilo e a estrutura do texto, o número de relações que se desejar extrair, a existência ou não de anotações no texto, se a informação se refere a um domínio específico ou é independente de domínio, entre outros.

Neste trabalho, os tipos de textos trabalhados se referem a um domínio específico (agrotóxicos). As características (estilo) encontradas nos textos foram: (i) tratavam-se de

textos técnicos; (ii) os textos apresentavam seções padronizadas, bem delimitadas, e com separação por tópicos/assuntos bem definidos; (iii) as sentenças não apresentavam com frequência uma estrutura do tipo sujeito/verbo/objeto, não tendo assim relações interessantes que pudessem ser extraídas; (iv) não possuíam anotações nem bases de conhecimento associadas. O processo geral de nossa abordagem pode ser ilustrado pela Figura 1.

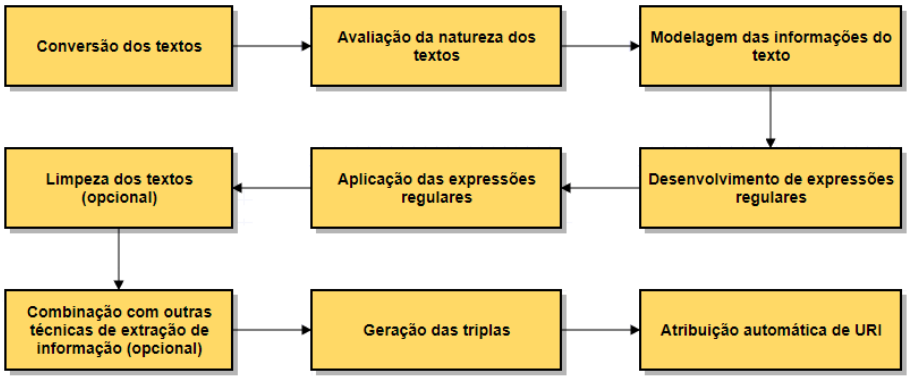

Figura 1. Abordagem geral de extração

O processo inicia com a conversão de qualquer formato de texto para o formato de texto simples (arquivos txt), visando facilitar o tratamento posterior. Em seguida, é feita uma avaliação da natureza do texto, verificando-se a ocorrência ou não de alguma estrutura ou padrão na representação de seu conteúdo. Esta avaliação permite que seja feita uma modelagem das informações contidas no texto, cujo objetivo é identificar o que há de interesse a ser extraído.

Em seguida, um conjunto de expressões regulares (ou outro método que se mostre mais adequado para a tarefa de extração) é definido e aplicado aos textos de interesse. Cabe notar que, em textos que são semi-estruturados, como por exemplo onde há a ocorrência de um padrão do tipo chave-valor, as expressões regulares podem ser definidas facilmente, apenas alterando-se o valor de interesse no padrão básico. As expressões regulares permitem que as informações extraídas já estejam no formato de triplas, ou quando isso não for diretamente possível, permitem selecionar as sentenças consideradas relevantes e que devem ser consideradas para passar por um processo de extração de informação.

Neste último caso, o conjunto de sentenças obtidas pelas expressões regulares necessita sofrer uma limpeza, com o objetivo de remover frases com sentido negativo e algumas stopwords. A partir deste conjunto de sentenças, é necessário combinar outras técnicas de extração de informação para a geração de triplas. Neste caso é gerada uma matriz TF-IDF que relaciona termos com documentos e indica a importância de um termo de um documento em relação a um conjunto de documentos. A matriz TF-IDF gerada é usada para gerar um conjunto de triplas, que relaciona um termo com os documentos nos quais tal termo é mais relevante.

Por fim, é feita a atribuição automática de URIs aos termos das triplas geradas, permitindo que os mesmos possam ser relacionados a outras bases e vocabulários existentes. Na próxima seção, mostraremos como esta abordagem foi aplicada em nosso

domínio específico, e apresentaremos então os experimentos e resultados obtidos.

4. Experimentos e Resultados

O domínio de agrotóxicos foi escolhido para aplicar a abordagem proposta na seção anterior. O objetivo é enriquecer semanticamente o portal do Observatório de Atenção Permanente ao Uso de Agrotóxicos, que reúne diversas fontes de dados disponibilizadas por diferentes instituições. O portal possui a vantagem de centralizar em um único ambiente virtual, de fácil usabilidade, as informações sobre agrotóxicos e seu consumo no Brasil.

Para alcançar este objetivo, estudamos os conjuntos de dados, tipos de documentos e material disponibilizados na *web* relacionados ao uso de agrotóxico no Brasil. Duas fontes de dados não estruturados chamaram atenção quanto à riqueza do seu contéudo: bulas encontradas no portal de agrotóxicos do Paraná [3] e monografias disponibilizadas no portal da ANVISA [4] sobre este tema. Para cada uma dessas fontes descreve-se a seguir a aplicação da abordagem tal qual apresentada na seção 3.

4.1. Aplicação da abordagem nas monografias

As monografias foram convertidas de textos em PDF para TXT e a conversão foi feita utilizando a biblioteca em Python chamada PyPDF2 [5]. Apesar das monografias estarem em formato textual, possuíam estilo semelhante e eram semi-estruturadas quanto ao seu conteúdo. Além disso, a parte essencial de informação de cada monografia era representada por uma estrutura ”chave:valor”, onde a chave era representada por propriedades referentes ao agrotóxico e o valor referente ao valor da propriedade.

As monografias continham as seguintes informações referentes ao agrotóxico: Ingrediente ativo ou nome comum; Sinonímia; Número CAS, que é o registro único de uma substância no banco de dados Chemical Abstract Service - CAS [6]; Nome Químico; Fórmula Bruta e Estrutural; Grupo Químico; Classe; Classificação Toxicológica; Uso Agrícola, que contém informações da modalidade de emprego daquele agrotóxico em diferentes culturas; e Ingestão Diária Aceitável.

A partir dessa avaliação da natureza dos textos, a modelagem das monografias foi realizada e é mostrada na Figura 2. Exemplos de leitura desta modelagem pode ser feita da seguinte maneira, um agrotóxico pode possuir 1 ou mais classes, e uma classe possuir 1 ou mais agrotóxicos, assim como um agrotóxico só pode possuir um Número CAS e um número CAS só pode estar associado a 1 agrotóxico.

Como as informações eram disponibilizadas da mesma forma e o domínio a ser aplicado era específico, então o uso de expressões regulares para extrair informações foi utilizado. Expressões regulares foram desenvolvidas para extrair e triplificar as informações modeladas, onde o *sujeito* era sempre o valor da propriedade *ingrediente ativo* e os *predicados* e *objetos* eram as *chaves* e *valores*, respectivamente. Em seguida, as expressões regulares foram aplicadas nas monografias e o resultado obtido foi um conjunto de triplas. Todas as triplas de cada monografia foram inseridas em um único arquivo

[3] http://celepar07web.pr.gov.br/agrotoxicos/bulas.asp

[4] http://portal.anvisa.gov.br/registros-e-autorizacoes/agrotoxicos/produtos/monografia-de-agrotoxicos/

[5] https://pypi.python.org/pypi/PyPDF2

[6] https://www.cas.org/

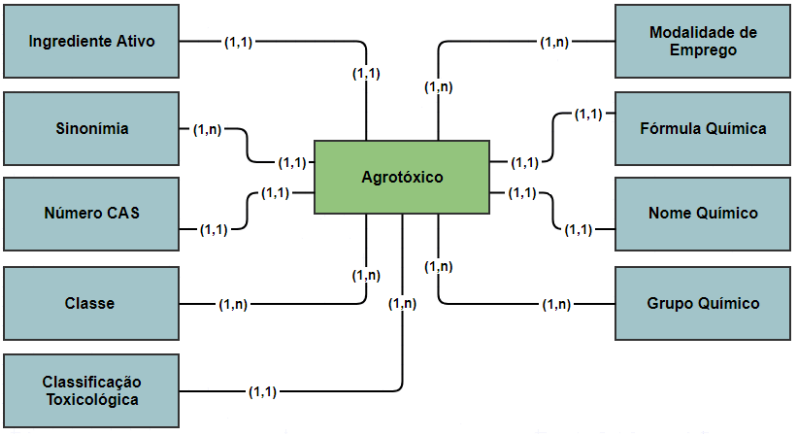

Figura 2. Modelagem de informações das monografias

CSV. Cada tripla foi inserida em uma linha e seus elementos separados por colunas. Depois de gerar o arquivo CSV, foi realizado uma limpeza para levar em consideração apenas as triplas que estivessem distribuídas em 3 colunas, retirando qualquer ruído para geração de triplas. Após a geração das triplas em formato RDF, a atribuição automática de URI foi realizada para cada *sujeito* distinto do CSV realizando busca automática do termo no dataset do Agrovoc [Caracciolo et al. 2013]. Caso o termo fosse encontrado, então a relação de igualdade era definida. Caso contrário, um vocabulário próprio era definido.

Para este experimento, foram utilizadas 475 monografias. Cada monografia corresponde a um agrotóxico (ou seja, temos 475 agrotóxicos). Destes, 147 foram encontrados no Agrovoc, onde cada agrotóxico encontrado recebeu a tripla de igualdade, como por exemplo (http://lodbr.ufrj.br/agrotoxicos/ACEFATO, owl:sameAs, http://aims.fao.org/aos/agrovoc/c_31235). Além disso, na própria geração das expressões regulares, as propriedades das triplas utilizaram vocabulários já existentes.

No total foram geradas 4644 triplas de acordo com as informações modeladas. Alguns exemplos de triplas geradas estão representadas no grafo da Figura 3.

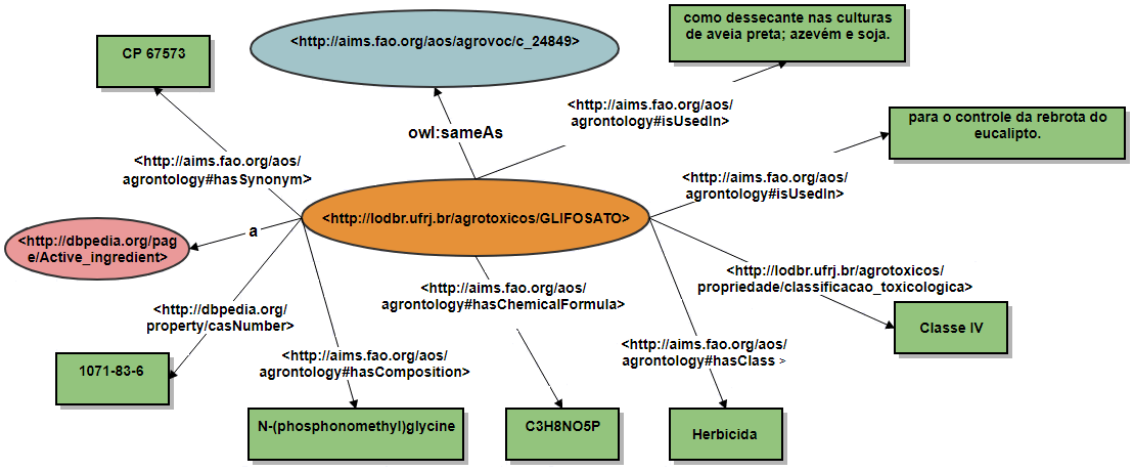

Figura 3. Exemplo de triplas geradas de uma monografia

4.2. Aplicação da abordagem nas bulas

As bulas também foram convertidas de PDF para TXT utilizando a mesma biblioteca citada acima. Apesar das bulas estarem em formato textual, possuíam um padrão na disponibilização de seu conteúdo, com seções bem definidas e delimitadas.

Uma bula contém as seguintes informações referentes a um agrotóxico: Classe; Composição; Tipo de Formulação; Classificação Toxicológica; Classificação do Potencial de Periculosidade; Primeiros Socorros; Número, Época e Intervalo de Aplicação; Dados Relativos à Proteção da Saúde Humana; Dados Relativos ao Meio Ambiente; Fabricante do Produto; Titular de Registro; Instrução de Uso (normalmente em tabela); Modo de Aplicação (Terrestre e Aérea); Procedimentos de Lavagem, Armazenamento, Transporte e Destinação de Embalagens Vazias e Restos de Produtos Impróprios para Utilização ou em Desuso.

A partir da avaliação da natureza do texto da bula, foi realizada uma modelagem. Nesta, foram desconsideradas informações que não tinham relação com o domínio do uso do agrotóxico e riscos a saúde. O resultado da modelagem das bulas é mostrado na Figura 4. Note que as duas fontes de dados (monografia e bula) puderam ser relacionadas através das informações relativas a Classe, Composição e Grupo Químico. Exemplos de interpretação desta modelagem pode ser feita da seguinte maneira: uma bula pode ou não ter efeitos colaterais, mas um efeito colateral pode estar associado a uma ou mais bulas, assim como uma bula possui apenas uma composição e uma composição só pode estar associada a uma bula.

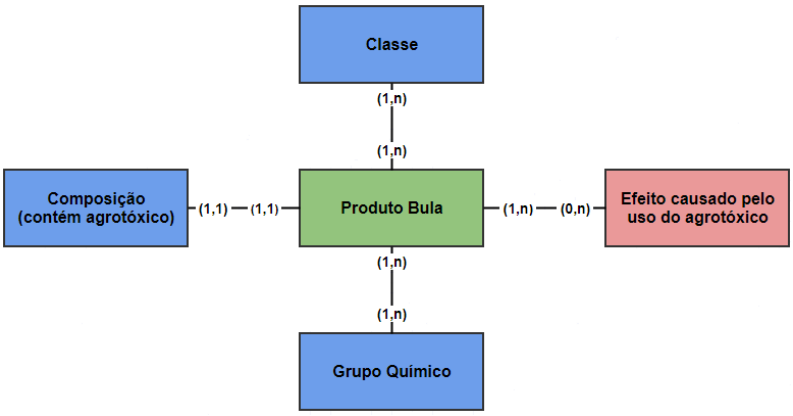

Figura 4. Modelagem de informações das bulas

Por conta da disposição das informações quase sempre ocorrer da mesma forma, expressões regulares foram desenvolvidas e aplicadas para selecionar apenas as seções modeladas como de interesse, pois a utilização apenas de expressões regulares não se mostrou suficiente para uma extração de qualidade. Por exemplo, determinar quais agrotóxicos (bulas) poderiam causar determinado efeito colateral. Isso ocorria porque parte da informação que estava descrita nas bulas era de texto não estruturado. Para extrair este tipo de informação, era interessante combinar outra técnica de extração de informação, como por exemplo uma matriz TF-IDF, onde o conjunto de termos de todas as bulas eram relacionados às bulas.

Antes de combinar esta técnica, para melhorar ainda mais o processo de extração

foi realizado um processamento para limpeza e preparação do texto. Como em alguns textos existiam frases com sentido negativo, ao se trabalhar com termos e não com a análise da frase completa, poderiam ocorrer casos de falsos positivos. Por exemplo: em uma bula com a frase ”não há evidências de câncer”, ao realizar a verificação dos termos, essa bula seria retornada para o termo ”câncer”, quando, na verdade, não deveria ser retornada. Desta forma, foram retiradas frases com sentido negativo. Um outro ponto tratado foi unificar termos em expressões com sentido próprio, como por exemplo ”aumento de pressão”, além de remover stopwords.

Com os textos processados, o próximo passo foi gerar uma matriz TF-IDF, cujo objetivo era extrair a relação de termos e documentos, onde o conjunto de termos de todas as bulas eram relacionados às bulas. Assim, identificando um sintoma dentro da lista de termos (por exemplo, vômito), foi possível relacionar este sintoma a todos os agrotóxicos (através de sua respectiva bula) que o causam. Com isso, um conjunto de triplas da forma (agrotóxico, provoca, sintoma) pode ser extraída.

Por último, foi realizada a atribuição automática de URIs para os termos da matriz TF-IDF. Por faltarem vocabulários e datasets em português sobre doenças e efeitos na saúde, os termos da matriz foram traduzidos para inglês e a partir dos termos em inglês foi realizada uma busca automática dos termos no dataset do BioPortal [Noy et al. 2009]. Caso o termo fosse encontrado, então a relação de igualdade era mapeada, e, em caso contrário, vocabulários próprios eram definidos.

Foram utilizadas 1234 bulas, separadas pelas categorias: *fungicida*, *herbicida*, *inseticida* e outros. O processo da bula foi executado nas quatro categorias. A Tabela 1 apresenta os resultados do processo.

Tabela 1. Resultados das bulas

Categoria	Documentos	Termos	Termos encontrados	Triplas
Inseticida	436	4186	369	2309
Herbicida	426	3539	261	1656
Fungicida	298	3093	235	1197
Outros	74	1173	61	140

Na categoria de inseticida, 436 documentos foram processados e 4186 termos foram gerados. O BioPortal foi utilizado por disponibilizar diversos vocabulários. Como nosso objetivo era apenas obter termos relacionados a sintomas e efeitos, restingimos a busca (via API) utilizando apenas a ontologia MEDDRA [7]. Além disso, utilizamos um POS Tagger do pacote NLTK [8] de modo a restringir nossa busca a termos que fossem substantivos. Após fazer as restrições, 369 termos foram encontrados e anotados, gerando 2309 triplas que associavam os sintomas às bulas. O mesmo processo foi realizado nas outras três categorias (ver Tabela 1).

Alguns exemplos de triplas geradas estão representadas no grafo da Figura 5.

[7] https://bioportal.bioontology.org/ontologies/MEDDRA
[8] https://www.nltk.org/

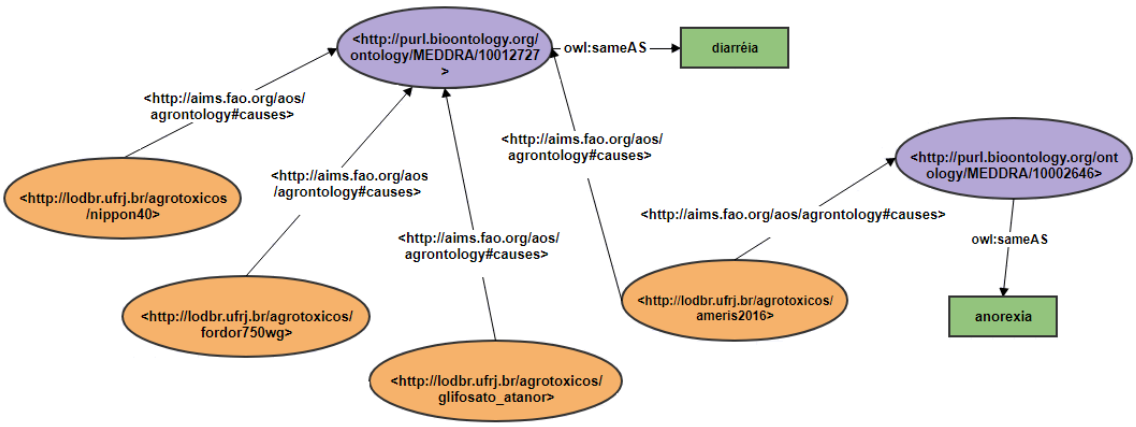

Figura 5. Exemplo de triplas geradas das bulas

4.3. Incorporação das triplas no portal

As triplas extraídas dos documentos estão sendo inseridas no portal do Observatório, para gerar o enriquecimento do mesmo e de outras bases que o mesmo referencia. No portal, outros dados estruturados, como CSV e XLSX, estão sendo triplificados, como por exemplo, um dataset que relaciona o agrotóxico e países onde o mesmo está proibido. Esses dados triplificados podem ser associados às triplas extraídas das monografias e bulas sendo, consequentemente, enriquecidos.

Além das informações dos sintomas extraídas das bulas, termos referentes à composição química, grupo químico e classe estão sendo associados com outros vocabulários e ontologias do BioPortal. Ao realizar esta associação o enriquecimento entre fontes de dados, monografias e bulas, também ocorrerá. Assim, por exemplo, a bula associada ao *glifosato atanor*, apresentada na Figura 5, possui em sua composição o glifosato. Este é o ingrediente ativo de um agrotóxico descrito em uma monografia, representado pela Figura 3, sendo possível verificar que o produto glifosato usado como dessecante nas culturas de aveia preta, azevém e soja pode ter como sintoma a diarréia.

5. Conclusão

Este trabalho apresentou uma abordagem para extração de dados e geração de triplas a partir de textos com características específicas. Com a ligação e associação das triplas com bases existentes e vocabulários externos, o portal do Observatório, que já vem tendo boa parte de seus datasets triplificados, passa a ter esses dados enriquecidos.

Apesar da abordagem ter sido experimentada em domínio específico, a ideia geral da mesma pode ser aplicada em outros domínios que possuírem estrutura/estilo de texto semelhante. A vantagem de usar essa abordagem é que a extração ocorre da forma desejada, não necessita de um corpus anotado, nem de bases de conhecimentos. A desvantagem é que para qualquer nova regra, a mesma deve ser adicionada na etapa de desenvolvimento das expressões regulares.

Este trabalho também abre novas possibilidades de trabalhos futuros, como por exemplo o enriquecimento de descritores de dados, e não só dados. Em diversos portais são utilizados conjuntos de tags para classificar e agrupar diferentes datasets que possuem algum conceito em comum. O problema é que as tags normalmente são atribuídas

por gestores de datasets de forma livre, estando sujeitas à ambiguidade e subjetividade. Técnicas já foram desenvolvidas para limpar, conciliar e enriquecer tags de portais de dados abertos de governo, como por exemplo em [Tygel et al. 2016], porém uma forma que parece interessante é tentar enriquecer semanticamente tags a partir dos dados e de extrações de fontes diversas.

Referências

Agichtein, E. and Gravano, L. (2000). Snowball: Extracting Relations from Large Plain-Text Collections. In *Proc. of the Fifth ACM Conf. on Digital Libraries*, DL '00, pages 85–94, New York, NY, USA. ACM.

Augenstein, I., Padó, S., and Rudolph, S. (2012). LODifier: Generating Linked Data from Unstructured Text. In *Proc. of the 9th Inter. Conf. on The Semantic Web: Research and Applications*, ESWC'12, pages 210–224, Berlin, Heidelberg. Springer-Verlag.

Brin, S. (1999). Extracting Patterns and Relations from the World Wide Web. In *Selected Papers from the Int. Workshop on The World Wide Web and Databases*, WebDB '98, pages 172–183, London, UK, UK. Springer-Verlag.

Byrne, K. and Klein, E. (2010). Automatic Extraction of Archaeological Events from Text. In *Proc. of the 37th Int. Conf. Computer App. and Quantitative Methods in Archaeology*, pages 48–56, Williamsburg, Virginia, USA.

Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., and Keizer, J. (2013). The AGROVOC Linked Dataset. volume 4, pages 341–348, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Carvalho, D. S., Freitas, A., and da Silva, J. C. P. (2013). Graphia: Extracting Contextual Relation Graphs from Text. In *The Semantic Web: ESWC 2013 Satellite Events - ESWC 2013 Satellite Events, Montpellier, France, May 26-30, 2013, Revised Selected Papers*, pages 236–241. Springer.

de Abreu, S. C., Bonamigo, T. L., and Vieira, R. (2013). A review on Relation Extraction with an eye on Portuguese. *Journal of the Brazilian Computer Society*, 19(4):553–571.

de Souza, E. N. P. and Claro, D. B. (2014). Extração de Relações utilizando Features Diferenciadas para Português. *Linguamática*, 6:57–65.

Del Corro, L. and Gemulla, R. (2013). ClausIE: Clause-Based Open Information Extraction. In *Pro. of the 22nd Int. Conf. on World Wide Web*, WWW '13, pages 355–366, New York, NY, USA. ACM.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open Information Extraction from the Web. *Commun. ACM*, 51(12):68–74.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artif. Intell.*, 165(1):91–134.

Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying Relations for Open Information Extraction. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1535–1545, Stroudsburg, PA, USA. Ass. for Comp. Linguistics.

Gamallo, P., Garcia, M., and Fernández-Lanza, S. (2012). Dependency-Based Open Information Extraction. In *Proc. of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, ROBUS-UNSUP '12, pages 10–18. Ass. for Comp. Linguistics.

Grishman, R., Sterling, J., and Macleod, C. (1991). Description of the Proteus System as used for MUC-3. In *Proc. of the Third Message Understanding Conference, San Diego, CA, May 1991*, pages 183–190. Morgan Kaufmann.

Joshi, A., Lal, R., Finin, T., and Joshi, A. (2013). Extracting Cybersecurity Related Linked Data from Text. In *Proc. of the 7th IEEE Int. Conf. on Semantic Computing*, pages 252–259. IEEE Computer Society Press.

Lange, D., Böhm, C., and Naumann, F. (2010). Extracting Structured Information from Wikipedia Articles to Populate Infoboxes. In *Proc. of the 19th ACM Int. Conf. on Inf. and Knowledge Management*, CIKM '10, pages 1661–1664, New York, USA. ACM.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2015). DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6(2):167–195.

Lehnert, W., Williams, R., Cardie, C., Riloff, E., and Fisher, D. (1991). The CIRCUS System as Used in MUC-3. Technical report, Amherst, MA, USA.

Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Med. Inf. Ass.*, 18(5):544–551.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A. D., Chute, C. G., and Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(Web-Server-Issue):170–173.

Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. ACL-44, pages 113–120, Stroudsburg, PA, USA. Ass. for Comp. Linguistics.

Tygel, A., Auer, S., Debattista, J., Orlandi, F., and Campos, M. L. M. (2016). Towards Cleaning-Up Open Data Portals: A Metadata Reconciliation Approach. In *Tenth IEEE Inte. Conf. on Semantic Comp., ICSC 2016, Laguna Hills, CA, USA, 2016*, pages 71–78.

Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., and Soderland, S. (2007). Textrunner: Open information extraction on the web. In *Proc. of Human Language Technologies: The Annual Conf. of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, NAACL-Demonstrations '07, pages 25–26, Stroudsburg, PA, USA. Ass. for Comp. Linguistics.