Anotação semântica automática de Objetos de Aprendizagem textuais em português com o paradigma *Open IE*[*]

Leandro M. P. Sanches[1], Laécio A. Costa[1], Marlo Souza[1], Laís N. Salvador[1]

[1]Programa de Pós-graduação em Ciência da Computação (PGCOMP)
Universidade Federal da Bahia (UFBA) – Salvador, BA – Brasil

{leandrompsanches, laeciocosta}@gmail.com, {laisns,msouza1}@ufba.br

Abstract. *This paper presents a work in progress that proposes a domain-independent model for automatic semantic annotation of textual LO in Portuguese. This model differs from other semantic annotation works by the use of Relations Disambiguation and the Open IE paradigm. For our model validation, we implemented a prototype which was able to extract information from textual LO and convert it into semantic metadata.*

Resumo. *O presente trabalho em andamento propõe um novo modelo independente de domínio para anotação semântica automática de OA textual em português que, em relação a trabalhos correlatos, destaca-se pela utilização da Desambiguação de Relações e do paradigma Open IE. Para a validação desse modelo, foi implementado um protótipo capaz de extrair informações de OA textuais e convertê-las em metadados semânticos.*

1. Introdução

Os Objetos de Aprendizagem (OA) são considerados como uma das formas mais adequadas para o desenvolvimento de conteúdo para a Educação a Distância através da internet, também conhecida como *e-learning* [Gonçalves 2007]. Os OA são quaisquer recursos digitais que possam ser reutilizados para oferecer suporte à aprendizagem [Wiley 2002].

A quantidade de OA tem aumentado nos últimos anos, esse fato pode ser comprovado no âmbito nacional, uma vez que existem diversos repositórios que armazenam e disponibilizam OA, entre eles: o Portal do Professor[1] e o WebEduc[2]. Em contrapartida, segundo Tarouco e Schmitt (2010), os OA apresentam baixa organização no seu armazenamento, o que dificulta a sua pesquisa e recuperação.

De acordo com Gonçalves (2007), os recursos da Web Semântica fornecem as ferramentas necessárias para realizar a anotação semântica de OA por pessoas e por agentes computacionais, facilitando a sua busca e recuperação. Contudo, a anotação semântica manual de OA geralmente é uma tarefa dispendiosa. Em contrapartida, são poucas as técnicas disponíveis na literatura que realizam a anotação semântica automatizada de OA, sendo que, especificamente em língua portuguesa, a quantidade de técnicas é ainda menor.

[*]O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

[1]http://portaldoprofessor.mec.gov.br/

[2]http://webeduc.mec.gov.br/

DOI: 10.5753/cbie.sbie.2018.1916

Neste contexto, o presente trabalho propõe um modelo independente de domínio para anotação semântica automática de OA textuais em língua portuguesa através de ontologias de domínio. A independência de domínio do modelo aqui proposto é obtida com a utilização do paradigma de Extração Aberta de Informações (do inglês, *Open Information Extraction (Open IE)*) [Wu e Weld 2010] e da parametrização da ontologia de domínio.

Além disso, a maioria dos estudos encontrados na literatura que anotam de forma automatizada OA textuais em português não obtém informações diretamente do texto do OA, por exemplo, a proposta de [Behr et al. 2016]. Entre os que extraem informações diretamente do OA, como em [Santanchè 2007], a extração é condicionada ao processo de autoria de um novo objeto. Assim, o modelo aqui proposto se diferencia dos demais encontrados na literatura por anotar OA através de instâncias e propriedades ontológicas obtidas diretamente do texto dos OA e de modo independente do processo de autoria.

2. Modelo de anotação semântica

O modelo proposto de anotação semântica, apresentado na Figura 1, é dividido em quatro etapas e prevê a utilização de ontologias de domínio e do padrão *Learning Object Metadata* (LOM) [IEEE LTSC 2002].

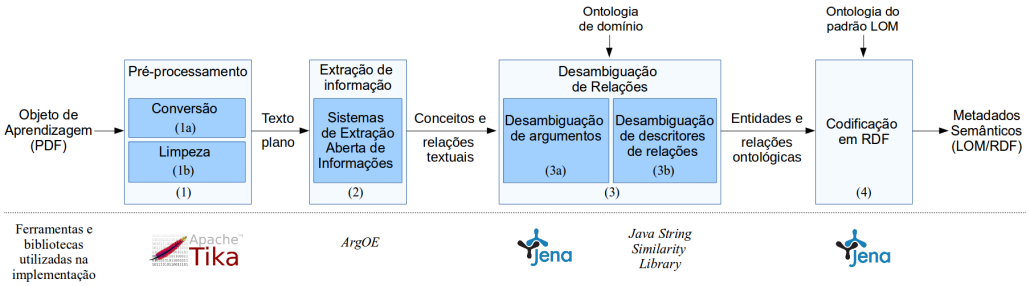

Figura 1. Etapas do modelo de anotação semântica

A primeira etapa do modelo de anotação objetiva extrair o conteúdo textual do OA do qual serão obtidos os metadados semânticos. Essa é iniciada com a conversão dos documentos originais para texto plano, atividade (1a) da Figura 1, e recebe como entrada um OA no formato *Portable Document Format* (PDF). Além do formato PDF, outros formatos podem ser considerados no futuro. Nessa etapa também é necessário realizar verificações e limpezas no texto para tratar possíveis problemas causados por incompatibilidades das diversas codificações de caracteres, atividade (1b) da Figura 1.

A segunda etapa, Extração de Informações, objetiva identificar os diferentes conceitos e as relações presentes no texto limpo do OA. Para isso, devem ser aplicadas técnicas de Processamento de Linguagem Natural (tokenização, segmentação de sentenças, etiquetagem morfossintática) e, devido ao caráter independente de domínio desse modelo, é utilizado um sistema de Extração Aberta de Informações. Esses sistemas são capazes de extrair, de modo independente de domínio, um grande conjunto de informações relacionais (arg_1, rel, arg_2). Sendo arg_i sintagmas nominais e o descritor de relação rel é um fragmento textual que indica uma relação semântica entre os argumentos [Wu e Weld 2010].

A terceira etapa, Desambiguação de Relações[3], utiliza uma ontologia de domínio desenvolvida em língua portuguesa. Essa etapa é responsável por associar os argumentos extraídos do texto com as entidades ontológicas, atividade (3a), e os descritores de relações com as propriedades da ontologia, atividade (3b), fornecendo assim a semântica e possibilitando a legibilidade por máquina dos dados extraídos.

Por fim, na última etapa, as relações semânticas desambiguadas são então codificadas como metadados semânticos no padrão LOM em linguagem *Resource Description Framework* (RDF), sendo que, para isso é utilizada uma ontologia descrevendo esse padrão. Caso necessário, os metadados semânticos gerados podem ser armazenados em repositórios, como por exemplo o *Apache Jena Fuseki*[4].

3. Avaliação do modelo de anotação semântica

A avaliação do modelo proposto foi dividida em duas partes. A primeira parte de avaliação contemplou a implementação de três métodos para a desambiguação dos descritores das relações. Esses métodos são baseados no cálculo da similaridade de cosseno com a utilização de modelos de agregação de palavras. A implementação e avaliação desses métodos foram descritas em [Sanches et al. 2018].

A segunda parte da avaliação contemplou a implementação de um protótipo de anotação semântica em linguagem Java. Nessa avaliação foi usada a ontologia de Arte Contemporânea [Trillo 2005] e a ontologia do padrão LOM[5]. Na implementação, conforme mostrado na Figura 1, a biblioteca *Apache Tika* foi utilizada para converter o conteúdo dos OA, contemplando a atividade (1a). Após a conversão do texto, realizou-se a normalização dos caracteres acentuados e a remoção de caracteres não latinos.

Subsequente, dado o conteúdo textual limpo do OA é preciso extrair as relações explicitamente contidas nesse texto. O extrator *ArgOE* [Gamallo e Garcia 2015] foi o escolhido para a Extração de Informações, etapa (2) na Figura 1. A escolha do *ArgOE* ocorreu devido a esse ser o único, para língua portuguesa, encontrado de modo acessível e gratuito. A Figura 2 mostra algumas relações extraídas de um OA[6], nesta cada palavra é associada a sua respectiva etiqueta morfossintática, por exemplo: em *Ana&Teixiera_NOUN* tem-se a palavra Ana Teixeira anotada como nome.

a_DT artista_NOUN-H brasileira_ADJ Ana&Teixeira_NOUN	sentou-se_VERB-H em_PRP	uma_DT cadeira_NOUN-H
a_DT artista_NOUN-H brasileira_ADJ Ana&Teixeira_NOUN	sentou-se_VERB-H em_PRP	vias_NOUN-H públicas_ADJ de_PRP diferentes_ADJ cidades_NOUN com_PRP uma_DT cadeira_NOUN vazia_ADJ
a_DT artista_NOUN-H brasileira_ADJ Ana&Teixeira_NOUN	sentou-se_VERB-H ao_PRP	seu_ADJ lado_NOUN-H
con_NOUN-H	se_PRO centrar_VERB-H no_PRP	gesto_NOUN-H
con_NOUN-H	se_PRO centrar_VERB-H de_PRP	ouvir_VERB-H
nas_PRP histórias_NOUN-H	contadas_VERB-H nas_PRP	histórias_NOUN-H
Ana&Teixeira_NOUN-H	propõe_VERB-H	um_DT encontro_NOUN-H

Figura 2. Relações textuais extraídas de um OA

[3]A Desambiguação de Relação pode ser entendida como a função que associa instâncias relações semânticas r, obtidas de uma ontologia, com instâncias de relações $(arg1, rel, arg2)$, extraídas por sistemas Open IE, que descrevam as mesmas informações que r [Sanches et al. 2018].

[4]https://jena.apache.org/documentation/fuseki2/

[5]http://lov.okfn.org/dataset/lov/vocabs/lom

[6]http://www.poiesis.uff.br/p25/p25-artigos-3-daniele-pires-de-castro.pdf

As relações extraídas do texto do OA são então passadas para a etapa (3) de Desambiguação de Relações. É nessa etapa que a ontologia de domínio se faz necessária e a sua implementação utilizou a biblioteca *Apache Jena*[7], que permite realizar consultas a ontologias. Assim, para realizar a desambiguação dos descritores das relações , atividade (3b) da Figura 1, foram considerados os três métodos anteriormente descritos e avaliados em [Sanches et al. 2018]. Já para a desambiguação dos argumentos, atividade (3a), aplicou-se a medida da distância de Levenshtein normalizada [Levenshtein 1966] disponível na biblioteca *Java-String-Similarity* [8].

Por fim, as relações extraídas do texto do OA e desambiguadas com a ontologia de domínio são então convertidas em metadados semânticos no padrão LOM, etapa de Codificação (4) da Figura 1. Para isso, novamente foi utilizada a biblioteca *Apache Jena* e também foram usados uma ontologia do padrão LOM e o modelo de codificação em RDF de metadados LOM de [Rajabi et al. 2014].

Embora ainda sejam necessários aprimoramentos no algoritmo de desambiguação, o protótipo é capaz de realizar a conversão das informações extraídas em metadados semânticos. A Figura 3 mostra parte de um metadado semântico extraído de um OA sobre arte contemporânea[9].

Conforme mostrado na Figura 3, o OA é anotado como uma instância da classe *lom:LearningObject* da ontologia LOM. As relações desambiguadas são associadas a essa instância pela propriedade ontológica *lom:description*. Sendo que, a Figura 3 mostra um metadado semântico especificado como a relação: *a autora da obra de arte Escuto Histórias de Amor atua como uma artista plástica*. O formato do OA, propriedade *dc:format*, foi automaticamente anotado através da DBpedia. Além disso, como o OA anotado não possuía um identificador único, esse foi gerado automaticamente e corresponde a propriedade ontológica *lom:identifier*. Por fim, é importante ressaltar que embora tenha ocorrido a parametrização da ontologia de domínio, o protótipo ainda não suporta a substituição da ontologia LOM.

```
<lom:LearningObject rdf:about="http://exemplo.org/f16a2a4c-6015-48ae-a038-15b5752e0638">
  <lom:identifier>
    <lom:Identifier rdf:about="http://exemplo.org/f16a2a4c-6015-48ae-a038-15b5752e0638#identifier">
      <lom:identifierEntry>f16a2a4c-6015-48ae-a038-15b5752e0638</lom:identifierEntry>
      <lom:identifierCatalog>UUID</lom:identifierCatalog>
    </lom:Identifier>
  </lom:identifier>
  <dc:format>http://dbpedia.org/resource/.pdf</dc:format>
  <lom:description>
    <rdf:Description rdf:about="http://a.com/ontology#Escuto_histórias_de_amor">
      <art:autor_atuando_como rdf:resource="http://a.com/ontology#Artista_Plástico"/>
    </rdf:Description>
  </lom:description>
  <lom:technicalSize>228506</lom:technicalSize>
  …
</lom:LearningObject>
```

Figura 3. Metadado semântico extraído no padrão LOM em RDF

4. Considerações finais

Neste trabalho foi proposto um modelo automático e independente de domínio para anotação semântica de OA. Para a sua elaboração foram utilizados os padrões recomen-

[7]https://jena.apache.org/
[8]https://github.com/tdebatty/java-string-similarity
[9]http://www.poiesis.uff.br/p25/p25-artigos-3-daniele-pires-de-castro.pdf

dados para Web Semântica, o padrão de metadados LOM e ontologias de domínio. O diferencial desse modelo é a utilização do paradigma *Open IE*, com o intuito de se obter independência de domínio e de anotar relações ontológicas no lugar de apenas instâncias de entidades. Por fim, um protótipo foi implementado para avaliar o modelo aqui proposto. Como trabalho futuro imediato, será realizada uma terceira etapa de avaliação do modelo contemplando as avaliações empíricas do protótipo e a investigação da influência das diferentes ontologias de domínio no processo de anotação semântica. Também será avaliada a possibilidade da anotação de outros campos de metadados.

Referências

Behr, A., Primo, T., e Viccari, R. (2016). Towards educational metadata interoperability on semantic web. In *27th Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1026–1035, Uberlândia. Brazilian Computer Society (SBC).

Gamallo, P. e Garcia, M. (2015). Multilingual open information extraction. In *Progress in Artificial Intelligence*, pages 711–722, Cham. Springer International Publishing.

Gonçalves, V. M. B. (2007). *A Web Semântica no contexto educativo: um sistema para a recuperação de objectos de aprendizagem baseado nas tecnologias para a Web Semântica, para o e-learning e para os agentes*. PhD thesis, University of Porto, Porto.

IEEE Learning Technology Standards Committee - IEEE LTSC (2002). IEEE standard for learning object metadata. *IEEE Std. 1484.12.1-2002*, pages 1–40.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710.

Rajabi, E., Sicilia, M.-A., Ebner, H., Palmer, M., e Sanchez, S. (2014). Recommendation on exposing ieee lom as linked data 1.0 (second version). http://data.organic-edunet.eu/ODS_LOM2LD/ODS_SecondDraft.html.

Sanches, L. M. P., Cardel, V. S., Machado, L. S., Souza, M. V. S., e Salvador, L. N. (2018). Disambiguating Open IE: identifying semantic similarity in relation extraction by word embeddings. In *Computational Processing of the Portuguese Language*. Springer International Publishing.

Santanchè, A. (2007). Otimizando a anotação de objetos de aprendizagem através da semântica in loco. In *18th SBIE*, pages 452–461, São Paulo. SBC.

Tarouco, L. M. R. e Schmitt, M. A. R. (2010). Adaptação de metadados para repositórios de objetos de aprendizagem. *Revista RENOTE: Novas Tecnologias na Educação*, 8(2).

Trillo, C. D. P. (2005). Recuperação de vídeos indexados por conceitos. Master's thesis, Universidade de São Paulo, São Paulo.

Wiley, D. A. (2002). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxionomy. In *The Instructional Use of Learning Objects*. Agency for Instructional Technology and Association for Educational Communications and Technology.

Wu, F. e Weld, D. S. (2010). Open information extraction using wikipedia. In *48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, Stroudsburg. Association for Computational Linguistics.