Comparativo entre fontes de dados para anotação automática de videoaulas

Jairo Francisco de Souza[1], **Jorão Gomes Jr.**[2], **Eduardo Barrére**[1]

[1]Departamento de Ciência da Computação – (UFJF)
36.360-900 – Juiz de Fora – MG – Brasil

[2]Bacharelado em Ciência da Computação – Universidade Federal de Juiz de Fora (UFJF)
36.360-900 – Juiz de Fora – MG – Brasil

{jairo.souza, joraojunior, eduardo.barrere}@ice.ufjf.br

Abstract. *The volume of the educational video repositories has rapidly increased as the creation of new digital resources is facilited. However, such repositories need better indexing and searching engines. Users find it difficult to find useful information when educational videos has poor quality metadata. The proccess of adding new descriptors to videos is called annontation. In this work, we will discuss how the information extracted from educational videos influence the automatic semantic annotation task. We will present an comparative analysis among these information sources and we will demonstrate which information sources are most useful for annontation task.*

Resumo. *A facilidade de criação de conteúdos digitais tem permitido que repositórios de recursos educacionais possam crescer rapidamente, gerando necessidade de melhores mecanismos de indexação e busca. Para usuários, ainda há muita dificuldade de encontrar vídeos de interesse quando estes possuem metadados de baixa qualidade. O processo de atribuir novos descritores aos vídeos é chamado de anotação. Este trabalho tem como objetivo explorar as fontes de informação presentes em videoaulas que podem auxiliar no processo no anotação semântica. O artigo apresenta uma análise comparativa entre a qualidade dessas fontes de dados de forma a demonstrar quais fontes geram melhor resultado para o processo de anotação.*

1. Introdução

A facilidade de criação de conteúdos digitais tem permitido que repositórios de recursos educacionais possam crescer rapidamente, gerando necessidade de melhores mecanismos de indexação e busca desses conteúdos. De todas as mídias, o vídeo é a que tem o maior destaque quando o assunto é educação, seja pela sua capacidade de adaptação às diversas plataformas [de Oliveira et al. 2010] ou pela sua atratividade natural, unindo fatores auditivos e visuais numa única mídia [Medeiros and Pansanato 2015]. Por outro lado, para usuários dos repositórios de videoaulas, é quase impossível encontrar vídeos de interesse quando se utiliza apenas pequenos trechos de texto associado ao vídeo, como título e alguns metadados disponíveis nos repositórios que, quase sempre, são curtos e de alto nível [Yang and Meinel 2014].

O fato dos usuários ainda não conseguirem extrair o máximo potencial em suas buscas mostra a relevância desses estudos, principalmente quando estão relacionados a

DOI: 10.5753/cbie.sbie.2017.1127

vídeos, onde o conteúdo dificilmente é bem representado semanticamente e a linguagem natural muitas vezes é vaga [Gupta et al. 2015]. Se tratando dos vídeos disponíveis na internet, isso se torna ainda mais evidente, pois existem grandes demandas para solucionar esse problema, atraindo muito interesse entre os pesquisadores [Jiang et al. 2013]. É comum encontrar grandes repositórios de videoaulas pouco descritas ou mal anotadas. Por exemplo, este artigo apresenta uma avaliação do serviço Videoaula@RNP e verifica-se que o serviço possui atualmente 858 vídeos. Destes, 604 estão mal descritos ou com anotações ruins (termos redundantes ou muito genéricos). No serviço da RNP e em outros repositórios públicos desse tipo, a descrição e anotação dos vídeos é de responsabilidade do produtor do vídeo e este nem sempre se preocupa com os metadados que está fornecendo durante o processo de submissão. Portanto, a aplicação de técnicas automáticas de anotação se tornam relevantes para garantir melhor acesso ao conteúdo digital.

Atualmente, diversas técnicas podem ser utilizadas para indexação de vídeos, como classificação por histogramas de cores, formas, reconhecimento de ações, reconhecimento de faces, extração de texto através de OCR (*Optical Character Recognition*), entre outras. Independente da fonte dos dados, os vídeos podem ser associados a conceitos, os quais expressam os principais assuntos abordados no vídeo. Este processo leva o nome de anotação semântica. Vídeos anotados semanticamente são relacionados a entidades que fazem parte de uma rede de relações semânticas, como uma ontologia ou um *thesaurus*. Mídias anotadas facilitam o processo de busca em repositórios e muitos pesquisadores têm trabalhado em melhorias no processo de anotação para diversas mídias [Qazi and Goudar 2016].

Este trabalho tem como objetivo explorar as fontes de informação presentes em videoaulas que podem auxiliar no processo de anotação semântica. O artigo apresenta uma análise comparativa entre a qualidade dessas fontes de forma a demonstrar quais geram melhor resultado para o processo de anotação. Embora existam diferentes estilos de vídeos educativos, é frequente o uso de vídeos reproduzindo aulas num formato mais próximo do tradicional, onde temos um professor (visível ou não) discursando sobre um conteúdo e fazendo uso de textos para descrever exemplos e gráficos exibidos durante a aula. Neste tipo de vídeo, podem ser identificados fontes de dados passíveis de anotação, como o texto exibido ao longo do vídeo, metadados como título, resumo e palavras-chave; o discurso falado do professor, legendas, etc. Assim, considerando alguns padrões de vídeos educativos existentes na web, este estudo explora estes tipos de dados mais encontrados em videoaulas e auxilia administradores de repositórios e pesquisadores a decidir por técnicas mais adequadas para indexação dos vídeos que disponibiliza, além de decidir quais são as fontes de informação que minimamente geram bons resultados.

Para melhor apresentação do trabalho, este está organizado da seguinte forma: a seção 2 apresenta uma discussão sobre anotação semântica, enquanto a seção 3 faz uma ligeira análise sobre os trabalhos relacionados. Em seguida, a metodologia utilizada nestes experimentos são descritas na seção 4 e os resultados são discutidos na seção 5. Por fim, a seção 6 apresenta as considerações finais.

2. Anotação semântica em vídeos

A busca em vídeos ainda é um problema difícil de se resolver dado ao fato da linguagem natural ser muitas vezes vaga e incerta, além de existir uma grande dificuldade de

representação semântica do seu conteúdo [Gupta et al. 2015]. Principalmente nos vídeos existentes na internet, soluções para esses problemas têm grandes demandas e assim acabam por atrair grande interesse entre os pesquisadores [Jiang et al. 2013].

Anotações semânticas podem ser feitas de forma manual ou automática. A anotação manual normalmente é feita uma única vez, por ser exaustivo anotar muitas cenas e falas. A forma automática é menos custosa e possui maior interesse da comunidade científica, visto que possui um custo muito menor e têm gerado resultados significativos.

Existem atualmente diversas soluções para busca em vídeos atribuindo a estes anotações semânticas que melhor descrevem seu conteúdo. As abordagens para atribuição dessas anotações podem ser divididas entre aquelas que fazem uso de dados externos relacionados ao vídeo e aquelas que utilizam apenas as informações contidas na mídia. No primeiro grupo, por exemplo, o uso de textos ao redor de imagens é utilizado por [Habibian et al. 2015] para verificar correspondências entre as imagens e textos e assim encontrar os conjuntos inter-relacionados de termos e temas, ao contrário de simplesmente anotar textos já definidos. Já [Maynard and Hare 2015] apresentam uma abordagem útil quando não há texto associado às mídias. Os autores abordam como a análise das questões sociais nas mídias podem ajudar os editores a selecionar material relevante e como a mineração de mídias sociais pode contribuir para relacionar arquivos. Eles propõem utilizar análise de sentimento e análise social para reconhecer se algumas mídias possuem significado importante, quando não é possível atribuir sentido somente aos meios textuais, por serem poucos ou irrelevantes. O uso de anotação manuais dentro de redes de colaboração também foram propostas para melhorar a atribuição automática de anotações e melhorar a busca em vídeos [Grünewald and Meinel 2015]. Em relação aos trabalhos que fazem uso apenas de informações contidas na mídia, podemos citar o trabalho de [Asghar et al. 2014], o qual apresentam algumas abordagens com conteúdo baseado em eventos e essas normalmente levam em consideração o conteúdo visual dos vídeos. Entre as abordagens apresentadas estão mecanismos de detecção de limite de tomadas, extração de quadros-chave para representação de partes importantes dos vídeos, estrutura de análise e segmentação de cena juntamente com técnicas de OCR para extração de recursos textuais e determinação de *tags*. Apesar dessas abordagens apresentarem bons resultados na classificação, elas são limitadas a tipos específicos de vídeos, normalmente, aqueles que apresentam uma estrutura parcial e temporal bem definidas.

Videoaulas, em sua grande maioria, são vídeos com conteúdo informativo uniforme, dificultando a identificação de eventos e a detecção de limite de tomadas. Estes tipos de vídeos não podem ser facilmente segmentados de acordo com um evento específico, cada parte do vídeo se mostra igualmente importante para o usuário [Taskiran et al. 2006]. Neste contexto, abordagens que fazem uso de reconhecimento de fala, como a apresentada em [Raimond and Lowis 2012] podem gerar informação útil para a anotação. [Gravier et al. 2015] afirmam que na maioria dos casos a fala, a linguagem e o áudio são importantes portadores de semântica nos conteúdos multimídia. Em particular, a linguagem é de extrema importância para a compreensão da natureza da mensagem. Ainda no cenário de vídeos informativos, o uso de OCR permite que a informação de slides apresentados durante o vídeo podem ser recuperados como fonte de anotação [Zhao et al. 2015].

3. Trabalhos relacionados

Embora existam diversas abordagens para anotação de vídeos, não são muitos os trabalhos que analisam a qualidade da informação utilizada para anotar semanticamente vídeos educacionais. Contudo, alguns pesquisadores avaliaram a qualidade da informação para o processo de indexação ou anotação de vídeos em outros contextos.

O uso de anotações semânticas para melhoria da busca em repositório de vídeo-aulas foi explorado em [Grünewald and Meinel 2015]. Segundo os autores, os trabalhos atuais em anotação manual de vídeos educacionais não levam em consideração a perspectiva do usuário e, então, estes realizaram experimentos com participantes de um MOOC para avaliar esta tarefa quando realizada de forma individual ou em grupo, a qualidade do uso de mapas de tópicos neste processo, e como a cultura da participação afeta a tarefa. Como resultados, os autores sustentam, dentre outras coisas, que o reuso de metadados pré-existentes no vídeo, como uso de OCR, pode auxiliar na tarefa e mostram que a participação de aprendizes neste processo auxilia no aprendizado.

Outros trabalhos exploraram a tarefa de indexação automática ou de anotação automática, como é o caso dos trabalhos de [Hauptmann et al. 2003], [Yang and Meinel 2014] e [Zhao et al. 2015]. Em [Hauptmann et al. 2003], os autores demonstram o efeito da extração e combinação de informação visual e de áudio para o processo de busca utilizando parte da Trilha de Recuperação de Vídeo do TREC 2001[1] (34 vídeos) como base de avaliação. Dentre as informações analisadas, estão o reconhecimento de fala, detecção de face, extração de texto via OCR e o uso de *image similarity matching*. Uma vez que a base do TREC é utilizada para avaliar abordagens de recuperação de informação, a entrada de cada teste é uma combinação de consulta textual, vídeos ou imagens de exemplos. Assim, para que as informações extraídas de cada vídeo sejam indexadas e relacionadas com a consulta, os autores realizaram os experimentos com dois modelos distintos: uma modelagem vetorial com similaridade de cossenos, simulando busca por texto, e um modelo probabilístico utilizado para recuperação de imagens. Ambas são técnicas conhecidas da literatura de recuperação de informação.

Em [Yang and Meinel 2014], por sua vez, apresentam uma abordagem para indexação e busca de vídeos educacionais. Os autores fazem uso de técnicas para segmentação automática dos vídeos e detecção de quadros-chave para montar um roteiro visual do vídeo e permitir navegação por conteúdo. No seu trabalho, experimentos são realizados para analisar o uso de OCR para extrair texto dos quadro-chaves que foram identificados e o uso de reconhecimento de fala nas faixas de áudio do vídeo. As técnicas utilizadas tem o objetivo de associar palavras-chaves aos vídeos. Os experimentos foram realizados com 12 alunos de mestrado que tinham que assistir vídeos de uma hora de duração e realizar buscas sobre o conteúdo dos vídeos. Os resultados mostraram que uma melhor acurácia foi alcançada quando os participantes faziam uso do conteúdo do vídeo e dos quadro-chaves, ao invés de usar apenas o sumário do vídeo, por exemplo.

Por fim, [Zhao et al. 2015] apresentam uma abordagem para melhoria na busca de vídeos educacionais da área biomédica através do uso de técnicas para navegação visual e textual do conteúdo do vídeo. Para montar a navegação do vídeo, são utilizadas técnicas para detecção de cenas, OCR e reconhecimento de fala para identificação de palavras-

[1] http://trec.nist.gov/

chaves para cada cena do vídeo. Foram realizados experimentos de acurácia da busca com 50 graduandos em uma base indexada de 25000 vídeos da área biomédica oriundos do YouTube. Embora os autores não tenham analisado o impacto do uso de cada tipo de metadado para o processo de indexação, foram apresentados diversos resultados de experimentos com diferentes ferramentas para apoiar cada etada do processo. Assim, os autores mostram que ferramentas para OCR e reconhecimento de fala apresentam, cada uma, uma acurácia em torno de 0,70 de medida F. Isto mostra que melhoria nas técnicas atuais de reconhecimento de fala e OCR podem gerar resultados ainda melhores para o processo de anotação semântica encontrados na seção 4.

A metodologia utilizada pelos trabalhos acima são semelhantes à metodologia utilizada neste trabalho. Contudo, os resultados e análises são distintas, visto que os trabalhos se diferem em relação à tarefa (anotações manuais em [Grünewald and Meinel 2015]; busca utilizando textos, imagens e vídeos de exemplo em [Hauptmann et al. 2003]; indexação por palavra-chave em [Yang and Meinel 2014]; construção de navegação visual de vídeos [Zhao et al. 2015]), à base de avaliação (*benchmark* do TREC em [Hauptmann et al. 2003]; base própria de avaliação em [Yang and Meinel 2014]; Youtube [Zhao et al. 2015]), aos critérios de avaliação ([Hauptmann et al. 2003] usa cobertura e ARR – *average reciprocal rank* – como métricas de avaliação, enquanto [Yang and Meinel 2014] usa medida F e [Zhao et al. 2015] e [Yang and Meinel 2014] avaliam o tempo que os participantes do experimento demoraram para terminar a tarefa de busca) e às variáveis de análise ([Hauptmann et al. 2003] usa detecção de face, OCR para recuperar nomes de pessoas e cargos – útil para trechos de documentários –; [Yang and Meinel 2014] usa detecção de quadros-chaves, OCR para criar automaticamente o sumário da aula, reconhecimento de fala para associar palavras-chaves ao vídeo; [Yang and Meinel 2014] analisa a acurácia das ferramentas existentes para OCR e reconhecimento de fala). Neste trabalho, é avaliado o impacto do uso de dados extraídos de vídeos educacionais para o processo de anotação semântica automática e a avaliação é realizada com uma base real de vídeos educacionais em língua portuguesa.

4. Processo de avaliação

Para realização dos experimentos, foi utilizada uma base de avaliação de vídeos educacionais em língua portuguesa (seção 4.1). Uma breve descrição sobre a forma de extração das fontes de dados utilizadas no experimento estão na seção 4.2. O impacto das fontes de dados foram avaliadas em duas abordagens de anotação semântica (seção 4.3).

4.1. Base de avaliação

Para avaliação da abordagem, foi criada uma base de avaliação com vídeos do serviço Videoaula@RNP[2] anotados manualmente. A base possui 39 vídeos em língua portuguesa das áreas de Ciência da Computação, Estatística, Química e Física. A base possui duração total de aproximadamente 6 horas de vídeo. Essas aulas foram assistidas por especialistas convidados a realizar o processo de anotação manual. Cada especialista atribuiu um recurso da DBpedia para cada assunto explicitamente falado durante o vídeo, sem repetição. Não houve restrição do número de recursos para cada vídeo que o especialista poderia atribuir. Durante o processo de criação da base, verificou-se o quão custoso é o processo

[2]http://videoaula.rnp.br/

de anotação manual. Para cada 1 hora de vídeo os especialistas demoraram em média 4 horas de trabalho manual, totalizando aproximadamente 24 horas para anotar toda a base. Toda a base está disponível[3] para uso por outros pesquisadores.

4.2. Extração dos dados das videoaulas

Informação textual de qualidade comumente é encontrada em videoaulas. Por exemplo, praticamente todas os vídeos do Videoaula@RNP foram gravados como uma aula expositiva, contendo projeção de slides ao longo do vídeo. Ainda, a maior parte do conteúdo informativo do vídeo está contida na fala do professor. Por isso, sistemas de busca que utilizam apenas metadados dos vídeos (título ou resumo) não conseguem atender o usuário quando este quer encontrar um vídeo utilizando termos que apareceram durante um exercício ou um exemplo que o professor abordou. Para avaliar as fontes de dados para anotação semântica, foram utilizadas:

- *Metadados:* foram extraídos os textos incluídos pelo criador do vídeo, o qual abrange o título do vídeo, o resumo e as palavras-chaves definidas pelo usuário e utilizadas pelo sistema de busca do Videoaula@RNP.
- *Sumário:* cada videoaula possui o sumário associado ao vídeo que descreve os tópicos que serão abordados ao longo da aula.
- *Reconhecimento de fala:* o áudio foi extraído e foi gerada a transcrição automática do áudio com o Kaldi[4] utilizando um modelo acústico e modelo de linguagem para português treinado pelos autores com videoaulas da RNP. A taxa de erro de palavras (WER) desse modelo é em torno de 40%. Embora o reconhecimento de fala possua diversos problemas, no contexto de vídeos educacionais tem-se um ambiente mais propício à sua utilização, uma vez que tem-se uma fala mais pausada em ambiente sem ruído, geralmente um único falante que está fazendo uso de equipamento adequado para captação do áudio.
- *Legenda:* a legenda manual contida nos vídeos substitui a transcrição automática. Ressalta-se que a legenda é uma transcrição manual adaptada para melhor se adequar à leitura e, assim, nem sempre possui exatamente o mesmo texto que foi falado. Essa adaptação, contudo, não tende a afetar o processo de anotação. Como a legenda nem sempre é encontrada em repositórios de vídeos por seu alto custo de produção, esta fonte de dado foi inserida nos experimentos para simular um processo de reconhecimento de fala com taxa de erro de palavras ótimo.
- *Reconhecimento de texto:* vídeos educativos geralmente possuem texto ao longo do vídeo ou arquivos PDF associados. Para utilizar essa informação quando não há PDF, o vídeo pode ser processado a cada *frame* com uma ferramenta de OCR. Os *frames* parecidos são descartados para que o texto gerado por cada *frame* seja único.

4.3. Abordagens para anotação semântica automática

Os experimentos foram realizados utilizando duas abordagens para anotação semântica: uma abordagem de *entity linking* e uma abordagem de extração de tópicos. Estas duas abordagens possuem objetivos distintos de anotação semântica e são influenciadas de

[3]https://github.com/—-removido para avaliação—-
[4]http://kaldi-asr.org

forma distinta de acordo com a qualidade do texto de entrada (como os gerados com certo grau de ruído oriundo dos processos de reconhecimento de fala e OCR).

A abordagem de reconhecimento de entidades possui como entrada um texto em linguagem natural e produz um conjunto de pares $(termo, entidade)$ que representam o conceito (entidade) associado ao termo presente no texto. Para esse processo, geralmente são utilizadas técnicas de processamento de linguagem natural para tokenizar o texto e identificar os termos corretos ("Rio de Janeiro" ao invés de "Rio" e "Janeiro"). Foi utilizado o DBpedia Spotlight [Mendes et al. 2011], o qual faz uso da DBpedia para criar um mapa de entidades candidatas a cada termo encontrado e desambiguar o termo, ou seja, associar a entidade que melhor descreve o termo. As entidades são identificadas para URIs da ontologia da DBpedia. Assim, os pares (*"Manchester mineira", dbpediaPT:Juiz_de_Fora*) e (*"industrialização", dbpediaPT:industrialização*) podem ser anotados à partir do texto *"Passou a ser conhecida como Manchester Mineira à época em que seu pioneirismo na industrialização a fez o município mais importante do estado"*.

Por sua vez, a extração de tópicos possui como entrada um texto em linguagem natural e produz um conjunto de entidades que representam os assuntos principais do texto. Este tipo de abordagem permite que entidades como dbpediaPT:Zona_da_Mata ou dbpediaPT:Cidades_de_Minas_Gerais sejam associadas ao texto, visto que possui grande relação ao conteúdo abordado pelo texto. Para esta tarefa, foi implementada a abordagem proposta em [Raimond and Lowis 2012], a qual faz uso do grafo de categorias da DBpedia para identificar as entidades com maior relação com o texto.

5. Resultados e discussão

Nos experimentos, foi realizada a anotação semântica automática com cada combinação de fonte de dado nas duas abordagens de anotação. Os resultados foram medidos utilizando a métrica de cobertura e a métrica de TopN foi utilizada como métrica de precisão. Seja um documento com um total de N_r anotações manuais não ranqueadas para um dado vídeo e que foram associados N_k anotações corretas pelo algoritmo. Seja $rank_i$ a posição da i-ésima anotação correta do conjunto de resposta, foi utilizada a equação abaixo. Uma constante $\alpha = 0.8$ foi adotada e esta ajusta a penalidade associada à posição da anotação correta na resposta. A métrica de TopN é utilizada para verificar não só se o algoritmo retornou resultados corretos, mas também o quão próximos estes resultados estão das primeiras posições. Os resultados dos experimentos estão na Tabela 1 e 2 para as abordagens de *entity linking* e extração de tópicos, respectivamente.

$$TopN = \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{\alpha^{rank_i}}{\sum_{j=1}^{rank_i} \alpha^j} \tag{1}$$

Em relação ao uso individual das fontes de dados, verifica-se as informação mais facilmente encontradas em videoaulas, como os metadados e o sumário, são também as que menos contribuem para a cobertura da anotação, o que mostra como os produtores de conteúdo tendem a associar texto de baixa qualidade para descrever o vídeo, o que dificulta a busca em sistemas que utilizam apenas essas informações. Em seguida, o reconhecimento de texto e o reconhecimento de fala permitem alcançar boa parte das anotações. Esses processos, como possuem uma taxa de erro no processo de reconhecimento, acabam influenciando na anotação. Vale notar que, para o processo de extração

Fonte	Cobertura	TopN
L	**0,838**	**0,020**
L+M	0,838	0,019
L+O	0,838	0,013
L+S	0,838	0,019
L+T	0,838	0,017
L+M+O	0,838	0,013
L+M+S	0,838	0,019
L+M+T	0,838	0,017
L+O+S	0,838	0,013
L+O+T	0,838	0,012
L+T+S	0,838	0,017
L+M+O+S	0,838	0,013
L+M+O+T	0,838	0,012
L+M+S+T	0,838	0,016
L+O+S+T	0,838	0,012
L+M+O+S+T	0,838	0,012
M	0,214	0,102
M+O	0,611	0,041
M+S	0,304	**0,102**
M+T	0,614	0,019
M+O+S	0,630	0,041
M+O+T	0,713	0,017
M+S+T	0,656	0,020
M+O+S+T	**0,720**	0,017
O	0,568	0,042
O+S	0,587	0,042
O+T	0,688	0,017
O+S+T	**0,694**	**0,017**
S	0,166	**0,175**
S+T	**0,590**	0,019
T	**0,531**	**0,017**

Tabela 1. *Entity Linking*

Fonte	Cobertura	TopN
L	0,387	0,156
L+M	0,410	0,171
L+O	0,432	0,303
L+S	0,387	0,194
L+T	0,387	0,151
L+M+O	0,432	0,334
L+M+S	0,410	0,185
L+M+T	0,387	0,162
L+O+S	0,432	0,362
L+O+T	0,454	0,327
L+S+T	0,387	0,172
L+M+O+S	0,432	**0,372**
L+M+O+T	0,454	0,337
L+M+S+T	0,387	0,172
L+O+S+T	0,454	0,344
L+M+O+S+T	**0,454**	**0,355**
M	0,071	0,076
M+O	0,286	0,132
M+S	0,132	0,118
M+T	0,287	0,160
M+O+S	0,291	0,129
M+O+T	0,351	0,165
M+S+T	0,305	0,161
M+O+S+T	**0,356**	**0,162**
O	0,281	0,109
O+S	0,285	0,120
O+T	0,347	0,145
O+S+T	**0,356**	**0,145**
S	0,098	0,098
S+T	**0,285**	**0,145**
T	**0,264**	**0,145**

Tabela 2. Extração de tópicos

L:Legenda, M: Metadados, O: OCR, S: Sumário, T: Transcrição

de tópicos, o uso de reconhecimento de texto ou de fala alcança um resultado bem mais próximo ao melhor resultado do que com a abordagem de *entity linking*. Isso se dá porque a primeira abordagem é menos dependente da linguagem natural do que a segunda e, assim, é menos afetada por ruídos gerados no texto. Isoladamente, a legenda foi a que gerou o melhor cobertura. Para vídeos educativos, este é um resultado esperado, uma vez que o locutor tende a repetir as informações que aparecem por escrito nos vídeos. Os resultados encontrados com o uso de legendas e reconhecimento de fala divergem dos apresentados por [Hauptmann et al. 2003], o qual argumenta que estas fontes não geram impacto relevante para a indexação, visto que a sua base de avaliação é mais diversificada e formada por vídeos promocionais, trechos de documentários, etc. Por outro lado, a legenda gera uma baixa precisão, principalmente na abordagem de *entity linking*, pois gera

um conjunto de texto muito grande e que muitas palavras que foram anotadas não estavam entre as entidades relevantes do vídeo. Neste caso, o uso do sumário ou metadados é mais adequado. Para a abordagem de extração de tópicos, legenda e transcrição, isoladamente, geram também boas precisões. Isso se dá porque a abordagem de extração de tópicos é influenciada pela frequência das palavras no texto. Assim, palavras que ocorrem repetidas vezes dentro do texto contribuem mais para que uma dada entidade seja atribuída ao texto.

Quando as fontes de dados são combinadas, é possível observar melhorias em ambas as abordagens de anotação. Para a abordagem de *entity linking*, a falta da legenda pode ser suprida com o uso das demais fontes de dados para uma melhor cobertura. Já para a abordagem de extração de tópicos, tem-se boa cobertura com o uso de O+S+T, sendo indiferente o uso de metadados. Ainda, verifica-se um melhor resultado em L+M+O+S+T. A legenda, neste caso, ajudou na cobertura e precisão por alterar a frequência dos termos e, também, por incluir palavras sem ruído no texto. Este é o caso também de alguns resultados podem ter aumentado, como a precisão do teste L+M+O+S em comparação ao teste L+M+O+S+T. Neste caso, a transcrição inclui palavras com ruídos, o que influencia negativamente na precisão. Contudo, verifica-se que as variações são quase desprezíveis.

6. Conclusões

Este trabalho mostrou como algumas fontes de dados presentes em vídeos educacionais podem influenciar o processo de anotação semântica automática. Foram utilizadas duas abordagens distintas de anotação para uma melhor compreensão do comportamento do uso dessas fontes de dados. Verificou-se que a legenda permite uma melhor cobertura dos anotações. Porém, por ser uma fonte de dado pouco frequente nos repositórios de videoaulas, dado o seu custo de produção, esta pode ser substituída pelo reconhecimento de áudio e demais fontes de dados, com bons resultados.

Como limitação desse estudo, podemos ressaltar o uso de bases de avaliação e a falta de uma análise de custo-benefício. Todo estudo que faz uso de bases de avaliação, estão sujeitos à qualidade dos dados de teste. Embora a base de avaliação utilizada tenha sido criada por especialistas, é possível que termos que tenham sido anotados corretamente por ambas abordagens não constem na base de avaliação. A base de avaliação foi criada para avaliar o quão similar o resultado de uma abordagem de anotação semântica está de anotações realizadas por seres humanos. Neste sentido, embora outras análises possam ser realizadas para verificar acurácia dos experimentos, a base de avaliação se mostra adequada para o estudo que foi proposto. Ainda, o estudo de custo-benefício de produção das fontes de dados ainda não foi totalmente explorado na literatura. Verificamos que legendas e reconhecimento de áudio são duas fontes de dados importantes para anotação de vídeos educacionais. Embora é de consenso na literatura que a produção de legendas é custosa, a produção de bons modelos para reconhecimento de áudio pode não é uma simplória. Assim, para repositórios com diferentes volumes, poderia-se verificar qual tipo de técnica gera melhores resultados para um menor custo.

Referências

Asghar, M. N., Hussain, F., and Manton, R. (2014). Video indexing: A survey. *International Journal of Computer and Information Technology*, 3(01).

de Oliveira, F. K., Santana, J. R., and de Oliveira Pontes, M. G. (2010). O vídeo como ferramenta educacional a partir de múltiplas plataformas. In *Brazilian Symposium*

on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE), volume 1.

Gravier, G., Jones, G. F., Larson, M., and Ordelman, R. (2015). Overview of the 2015 workshop on speech, language and audio in multimedia. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1347–1348. ACM.

Grünewald, F. and Meinel, C. (2015). Implementation and evaluation of digital e-lecture annotation in learning groups to foster active learning. *IEEE Transactions on Learning Technologies*, 8(3):286–298.

Gupta, Y., Saini, A., and Saxena, A. (2015). A new fuzzy logic based ranking function for efficient information retrieval system. *Expert Systems with Applications*, 42(3):1223–1234.

Habibian, A., Mensink, T., and Snoek, C. G. (2015). Discovering semantic vocabularies for cross-media retrieval. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 131–138. ACM.

Hauptmann, A. G., Jin, R., and Ng, T. D. (2003). Video retrieval using speech and image information. In *Electronic Imaging 2003*, pages 148–159. International Society for Optics and Photonics.

Jiang, Y.-G., Bhattacharya, S., Chang, S.-F., and Shah, M. (2013). High-level event recognition in unconstrained videos. *International journal of multimedia information retrieval*, 2(2):73–101.

Maynard, D. and Hare, J. (2015). Entity-based opinion mining from text and multimedia. In *Advances in Social Media Analysis*, pages 65–86. Springer.

Medeiros, S. F. d. L. and Pansanato, L. (2015). Estudo das preferências de alunos e professores sobre videoaula para identificar requisitos de software para ferramentas de produção. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 26, page 219.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM.

Qazi, A. and Goudar, R. (2016). Emerging trends in reducing semantic gap towards multimedia access: A comprehensive survey. *Indian Journal of Science and Technology*, 9(30).

Raimond, Y. and Lowis, C. (2012). Automated interlinking of speech radio archives. *Linked Data on the Web (LDOW'16)*.

Taskiran, C. M., Pizlo, Z., Amir, A., Ponceleon, D., and Delp, E. J. (2006). Automated video program summarization using speech transcripts. *IEEE Transactions on Multimedia*, 8(4):775–791.

Yang, H. and Meinel, C. (2014). Content based lecture video retrieval using speech and video text information. *IEEE Transactions on Learning Technologies*, 7(2):142–154.

Zhao, B., Xu, S., Lin, S., Luo, X., and Duan, L. (2015). A new visual navigation system for exploring biomedical open educational resource (oer) videos. *Journal of the American Medical Informatics Association*, 23(e1):e34–e41.