Exploiting Multimodality in Video Hyperlinking to Improve Target Diversity

Rémi Bois, Vedran Vukotić, Anca-Roxana Simon, Ronan Sicre, Christian Raymond, Pascale Sébillot, Guillaume Gravier

Exploiting multimodality in video hyperlinking to improve target diversity

Rémi Bois[1], Vedran Vukotić[2], Anca-Roxana Simon[4], Ronan Sicre[3], Christian Raymond[2], Pascale Sébillot[2], and Guillaume Gravier[1]

[1] CNRS, IRISA & Inria Rennes firstname.lastname@irisa.fr
[2] INSA Rennes, IRISA & Inria Rennes firstname.lastname@irisa.fr
[3] Inria, IRISA & Inria Rennes firstname.lastname@irisa.fr
[4] Univ. Rennes 1, IRISA & Inria Rennes anca-simon@outlook.com

Abstract. Video hyperlinking is the process of creating links within a collection of videos to help navigation and information seeking. Starting from a given set of video segments, called anchors, a set of related segments, called targets, must be provided. In past years, a number of content-based approaches have been proposed with good results obtained by searching for target segments that are very similar to the anchor in terms of content and information. Unfortunately, relevance has been obtained to the expense of diversity. In this paper, we study multimodal approaches and their ability to provide a set of diverse yet relevant targets. We compare two recently introduced cross-modal approaches, namely, deep auto-encoders and bimodal LDA, and experimentally show that both provide significantly more diverse targets than a state-of-the-art baseline. Bimodal autoencoders offer the best trade-off between relevance and diversity, with bimodal LDA exhibiting slightly more diverse targets at a lower precision.

1 Introduction

The automatic generation of hyperlinks within video collections recently became a major subject, in particular via evaluation benchmarks within MediaEval and TRECVid [9, 8, 20]. The key idea is to create hyperlinks between video segments within a collection, enriching a set of anchors that represent interesting entry points in the collection. Links can be seen as recommendations for potential viewers, whose intent is not known at the time of linking. The goal of the links is thus to help viewers gain insight on a potentially massive collection of videos so as to find information of interest, following a search and browse paradigm.

Creating video hyperlinks from a given anchor traditionally implements two steps. A segmentation step that aims at determining potential targets is followed by a selection step in which relevant targets are selected. The vast majority of approaches developed for the selection step rely on direct pairwise content-based similarity, seeking targets whose content is very similar to the anchor. Unsurprisingly, most use textual and/or visual content comparison [14, 7, 2, 17, 12, 1, 6, 21]. Maximizing content-based similarity between anchors and targets

showed to offer good relevance, as evidenced in [14] where n-gram bag-of-words are used to emphasize segments sharing common sequences of words.

Unfortunately, emphasizing relevance by rewarding highly similar content in terms of words and visual concepts does not offer diversity in the set of targets that are proposed for a given anchor. This lack of diversity is considered as detrimental in many exploration scenarios, in particular when users' intentions and information needs are not known at the time of linking. In this case, providing relevant links that cover a number of possible extensions with respect to the anchor's content is desirable. Clearly, having a set of diverse targets strongly improves the chance for any user to find at least one interesting link to follow, whatever his/her initial intentions. Additionally, target diversity directly improves serendipity, *i.e.*, unexpected yet relevant links, offering the possibility to drift from the initial anchor in terms of information so as to gain a better understanding of what can be found in the collection. This objective of providing diverse results is gaining traction in the hyperlinking community and has been explicitly included in the latest TRECVid evaluation benchmark [10].

In this paper, we investigate cross-modal approaches recently introduced for multimedia content matching, namely, bimodal autoencoders [27] and bimodal LDA [4], as a mean to improve the diversity of targets within a limited set. The intuition is that cross-modality unveils relevant links that would not be captured with standard approaches, and is a good candidate to improve diversity. Indeed, providing links to visual content related to spoken content, and conversely, is bound to reduce the similarity between anchors and targets while maintaining high relevance. For instance, a target could talk about what is shown in the anchor or show things that are discussed in the anchor, and thus bring complementary information. While multimodal approaches have been proposed to increase content similarity between anchors and targets [17, 1, 6], cross-modality has been seldom considered so far and no evaluation regarding the diversity of targets has been run to this date. In Sections 2 and 3, we introduce the two cross-modal systems that are used in this study, namely a bidirectional deep neural network, and a cross-modal topic model. Section 4 describes the evaluation protocol used to assess diversity, presents the corpus, and discusses results obtained by a user-centric study, as well as automatic measures.

2 Bidirectional Deep Neural Networks

The first approach that we consider to improve diversity relies on distributional representations of words and their multimodal extensions. Word vector representations, such as *word2vec* [19], have proven to be of interest for information retrieval [16, 18, 29], and were recently experimented for video hyperlinking [21, 27]. Interestingly, this representation of words can easily be used for cross-modal matching, often in conjunction with deep neural networks [30, 11, 27], with a strong potential for diversity. In this study, we rely on bidirectional symmetrical deep neural networks (BiDNN) operating on averaged *word2vec* representa-

tions [5] of words, obtained by automatic transcription, and on visual concepts detected in keyframes.

Autoencoders are neural networks, used in unsupervised learning, that are setup to reconstruct their input, while the middle layer is being used as a new representation of the data. In a multimodal setting, autoencoders are used to combine separate input representations to yield a joint multimodal representation in the middle hidden layer.

Typical multimodal autoencoders come in two varieties: i) extended classical single-modal autoencoders where multimodality is achieved by concatenating the modalities at their inputs and outputs and ii) truly multimodal autoencoders that have separate inputs and outputs for each modality, as well as one or more separated fully connected layers assigned to each. Both share a common central point: a fully connected layer connecting both modalities and used to obtain a multimodal embedding. However, multimodal autoencoders have some downsides. First, to enable cross-modal translation, one modality is often sporadically removed from the input, while autoencoders are asked to reproduce both modalities at their output. This means that an autoencoder has to learn to represent the same output both when a specific modality is present and when it is zeroed, which is less optimal than having direct cross-modal translation. Secondly, central fully connected layers are influenced by both modalities (either directly or through other fully connected layers). While this is good for multimodal embedding, it does not provide a clean cross-modal translation.

Bidirectional symmetrical deep neural networks tackle these two problems by first creating straight-forward cross-modal translations between modalities and then providing a common representation space where both modalities are projected and a multimodal embedding is formed. Learning is performed in both directions: one modality is presented as input and the other as the expected output while, at the same time, the second modality is presented as input and the first as expected output. This architecture is presented as two networks—one translating from the first modality to the second and the other conversely—where the variables in the central part are tied to enforce symmetry, as illustrated in Figure 1. Implementation-wise, the variables representing the weights in the hidden layers are shared across the two networks and are in fact the same variables. Learning the two cross-modal mappings is thus performed simultaneously thanks to the symmetric architecture in the middle. The joint representation formed in the middle layer while learning acts as a multimodal pivot representation enabling translation from one modality to the other.

Formally, let $\mathbf{h_i^{(j)}}$ denote (the activation of) the hidden layer at depth j in network i ($i = 1, 2$, one for each modality), $\mathbf{x_i}$ the feature vector for modality i and $\mathbf{y_i}$ the output of the network for modality i. Networks are defined by their weight matrices $\mathbf{W_i^{(j)}}$ and bias vectors $\mathbf{b_i^{(j)}}$, for each layer j, and admit f as activation function. The entire architecture is then defined by:

$$\mathbf{h_i^{(1)}} = f(\mathbf{W_i^{(1)}} \times \mathbf{x_i} + \mathbf{b_i^{(1)}}) \qquad i = 1, 2 \tag{1}$$

$$\mathbf{h_1^{(2)}} = f(\mathbf{W^{(2)}} \times \mathbf{h_1^{(1)}} + \mathbf{b_1^{(2)}}) \tag{2}$$

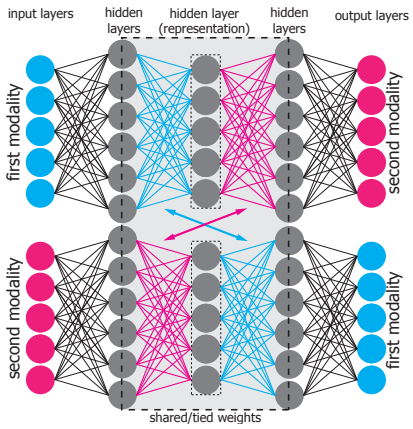

Fig. 1. Architecture of bidirectional symmetrical deep neural networks [27]

$$\mathbf{h_1^{(3)}} = f(\mathbf{W^{(3)}} \times \mathbf{h_1^{(2)}} + \mathbf{b_1^{(3)}}) \tag{3}$$

$$\mathbf{h_2^{(2)}} = f(\mathbf{W^{(3)^T}} \times \mathbf{h_2^{(1)}} + \mathbf{b_2^{(2)}}) \tag{4}$$

$$\mathbf{h_2^{(3)}} = f(\mathbf{W^{(2)^T}} \times \mathbf{h_2^{(2)}} + \mathbf{b_2^{(3)}}) \tag{5}$$

$$\mathbf{o_i} = f(\mathbf{W_i^{(4)}} \times \mathbf{h_i^{(3)}} + \mathbf{b_i^{(4)}}) \qquad i = 1, 2 \tag{6}$$

It is important to note that the weight matrices $\mathbf{W^{(2)}}$ and $\mathbf{W^{(3)}}$ are used twice due to weight tying, respectively in Eqs. 2, 5 and Eqs. 3, 5. Training is performed by applying batch gradient descent to minimize the mean squared error of $(\mathbf{o_1}, \mathbf{x_2})$ and $(\mathbf{o_2}, \mathbf{x_1})$ thus effectively minimizing the reconstruction error in both directions and creating a joint representation in the middle.

Given such an architecture, cross-modal translation can be done straightforwardly by presenting the first modality as $\mathbf{x_i}$ and obtaining the output in the representation space of the second modality as $\mathbf{y_i}$. However, to improve relevance while preserving diversity, we experimented with the multimodal embedding of the hidden layer. In practice, for a video segment, each modality is projected to the hidden layer with the corresponding network and the two resulting vectors are concatenated. More specifically, if both modalities are present, each one is presented to its respective input of the bidirectional deep neural network and the values are propagated through the network. The values from the central layer, where the common representation space lies, are concatenated and a multimodal embedding is formed. When only one modality is available, it is presented to its respective input of the bidirectional deep neural network and the values are propagated to the network. The values of the central layer are duplicated, as to form an embedding of an equal size as when both modalities are present. This allows for transparent comparison of video segments regardless of modality availability. When one video segment has only one modality while the other has both, a distance computed on such multimodal embedding would automatically compare the one available modality from one video segment with

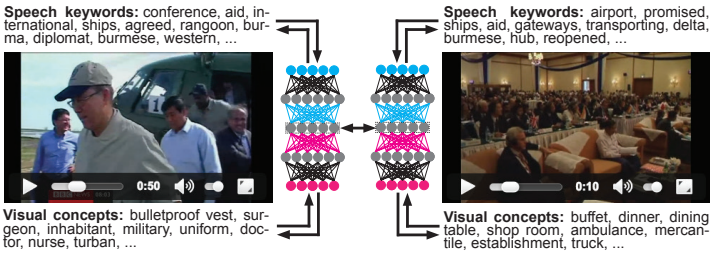

Fig. 2. Video hyperlinking with bidirectional symmetrical deep neural networks [28]

the two modalities of the other video segment. This all happens in the new common representation space where both modalities are projected and transparent comparisons are made possible.

Note that while the embedding is multimodal, it corresponds to a space dedicated to cross-modal matching and thus significantly differs from classical joint multimodal spaces. Figure 2 illustrates the task of video hyperlinking with bidirectional deep neural networks: for all video segments, cross-modal translations between embedded automatic transcripts, embedded visual concepts, and back, are learned. Then, for the specific video segments that are compared, their respective embedded automatic transcripts and embedded visual concepts are presented (regardless of modality availability) and their multimodal embeddings in the new common representation space are formed. Finally, the two multimodal embeddings are compared with a cosine distance to obtain a similarity score.

We implemented bidirectional neural networks in *Lasagne*[5]. All embeddings have a dimension of 100 as larger dimensions did not bring any significant improvement. The architectures used had 200-100-200 hidden layers as other smaller sizes performed worse and larger sizes did not perform better. The networks were trained with stochastic gradient descent (SGD) with Nesterov momentum, dropout of 20%, in mini-batches of 100 samples, for 1000 epochs (although convergence was achieved quite earlier). Since all the methods described belong to unsupervised learning, the learning was performed on the part of the dataset that contains both transcripts and visual concepts and tested on the whole dataset.

3 Cross-modal topic model

Another potential solution to diversity is the use of topic models, such as latent Dirichlet allocation (LDA), where the similarity between two documents is measured via the similarity of the latent topics they share rather than by direct content comparison [3]. Recently, based on seminal work on multilingual topic modeling [24], multimodal extensions of LDA were proposed for cross-modal video hyperlinking [4], combining the potential for diversity offered by topic models and by multimodality. As for BiDNN, words extracted from the

[5] https://github.com/Lasagne/Lasagne

Topic 3	words	love, home, feel, life, baby
	visual concepts	singer, microphone, sax, concert, flute
Topic 7	words	food, bit, chef, cook, kitchen
	visual concepts	fig, acorn, pumpkin, guava, zucchini
Topic 25	words	years, technology, computer, key, future
	visual concepts	tape-player, computer, equipment, machine, appliance

Table 1. Three multimodal topics represented by their top-5 words and visual concepts

automatic transcripts and the visual concepts from the keyframes are used in the bimodal LDA (BiLDA).

The LDA model is based on the idea that there exist latent variables, *i.e.*, topics, which explain how words in documents have been generated. Fitting such a generative model to a document means finding the best set of such latent variables in order to explain the observed data. As a result, documents are seen as mixtures of latent topics, while topics are probability distributions over words. The multimodal extension in [4] considers that each latent topic is defined by two probability distributions, one over each modality (or language in [24]). The BiLDA model is thus trained on parallel documents, assuming that the underlying topic distribution is common to the two modalities. In the case of videos, parallel documents are straightforwardly obtained by considering the transcripts and the visual concepts of a video segment as two parallel documents sharing the same underlying topic distribution. Training, *i.e.*, determining the topics from a given collection of videos, is achieved by Gibbs sampling, as for standard LDA [25], with the number of latent topics set to 700. Given a set of documents in the text (resp. visual) modality with vocabulary V_1 (resp. V_2), the probability that a word $w_i \in V_1$ (resp. visual concept $c_i \in V_2$) corresponds to topic z_j is estimated as

$$p(w_i|z_j) = \frac{n_{z_j}^{w_i} + \beta}{\sum_{x=1}^{|V_1|} n_{z_j}^{w_x} + \beta|V_1|} \quad , \tag{7}$$

where $n_{z_j}^{w_i}$ is the number of times that topic z_j was assigned to word w_i in the training data and β is a Dirichlet prior.

This training step provides a mapping between topics of the two modalities by means of the topics. Table 1 displays examples of this mapping obtained from the corpus used in this study (see corpus description Section 4). For each topic, we show the 5 most probable words and visual concepts. Sometimes, words and visual concepts are directly related (e.g., *computer* in topic 25). However, the relation can be more subtle, as in topic 3 where visual concepts describe a stage, and words are utterances frequently encountered in the lyrics of songs.

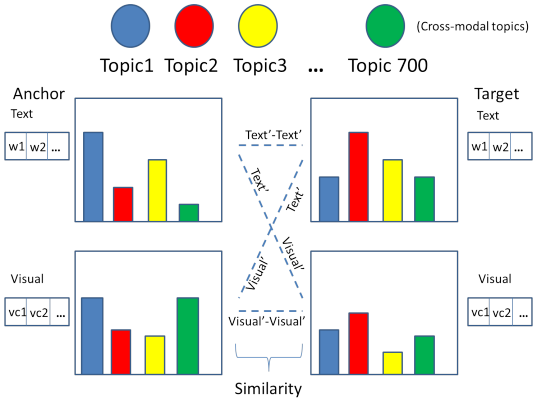

Fig. 3. Illustration of the multimodal and cross-modal matching with the BiLDA model

The interest of topic models lies in the fact that video segments dealing with similar topics will tend to have similar distribution over the latent topics. This enables the indirect comparison of two video segments by comparing the distribution of latent topics, rather than using their multimodal content, thus potentially enabling a diversity of content (within documents from closely related topics). Formally, given a video segment d, the idea is to represent the segment as a vector collecting the topic probabilities

$$p(d|z_j) = \left(\prod_{i=1}^{n_x} p(w_i|z_j) \right)^{1/n_x} , \tag{8}$$

where n_x is the size of the vocabulary in d and w_i is the i^{th} word or visual concept in d. Note that $p(d|z_j)$ is an approximation of the posterior $p(z_j|d)$, considering a uniform distribution of topics, which is a reasonable assumption. The similarity score between any two segments is given by a cosine similarity between their corresponding vectors after $L2$-normalization.

In practice, the probabilities $p(d|z_j)$ can be obtained from either one of the two modalities (using the corresponding distributions $p(\cdot|z_j)$), thus enabling multimodal and cross-modal matching as illustrated in Fig. 3. In this paper, we considered visual to text matching, representing the distribution of topics based on visual concepts for the anchor and on automatic transcripts for the targets.

4 Experimental results

4.1 Experimental setup

Experimental evaluation of the different methods and of the diversity of targets is performed using data from the TRECVid 2015 video hyperlinking task [20] and the corresponding annotations. The original data consists of approximately 2,700 hours of BBC programs from which 100 anchors were defined, with an

anchors	baseline	BiDNN	BiLDA
all	0.59	0.57	0.24
16	0.80	0.80	0.40

Table 2. Precision at rank 10 on target reranking

average length of 71 seconds. Anchors were selected by experts as being segments of interest that a user would like to know more about. As a result of the 2015 evaluation, a set of potential targets along with relevance judgments is also provided, these targets being the top-10 targets proposed by each participating team. Relevance assessment of each of those targets was achieved post hoc on Amazon Mechanical Turk (AMT). In total for the 100 anchors, 21,176 targets are available with their relevance judgments, of which 25.4 % are actually relevant.

In this work, the content matching methods are evaluated via a reranking task where the set of potential targets are reordered for each anchor, thus getting rid of segmentation issues. For each anchor, reranking operates on an average of 212 targets proposed in 2015. Apart from practical reasons due to the lack of extensive ground-truth on the whole data set, the reranking task is justified by the fact that we want to assess the properties of different methods for the target *selection* step. We should however stress a minor bias in this setting due to the fact that anchors were initially proposed by the 2015 participants. Hence, the targets that we rerank are all somehow related to the anchor and we cannot appreciate the potential of the methods to discard totally irrelevant targets. This, however, does not hinder the potential of the methods compared here, in particular with respect to diversity.

Anchor and target segments are described according to two modalities. On the one hand, automatically generated speech transcripts [13] provide a lexical representation after lemmatization and stopword removal. On the other hand, the automatic detection of 1,537 visual concepts [26], averaged over the keyframes of a segment, provides a visual representation. As some anchors are very short, a context of 30 seconds around each actual anchor is considered.

4.2 Results

We first compare BiLDA and BiDNN with a transcript-only baseline system, on their ability to find relevant targets. The baseline system implements a bag-of-words representation for each segment with tf-idf weighting [22] along with cosine similarity. Inverse document frequencies were estimated on the set of anchors plus the set of proposed targets. The BiDNN was implemented with an architecture of 200-100-200 hidden layers, dropout in the central part of 50 % and trained with batch gradient descent on the video segments that appear on the groundtruth. The BiLDA model was trained on the full TRECVid 2015 dataset.

Results are reported in Table 2, where precision at 10 are given for the whole set of anchors and for the 16 anchors that gave the best results, and which were retained for perceptual study of diversity (see below). Results for the baseline

	transcripts			concepts		
	n_u	$\overline{d}_a$	$\overline{d}_i$	n_u	$\overline{d}_a$	$\overline{d}_i$
baseline	29.8	0.51	0.61	35.6	0.61	0.71
BiDNN	40.8	0.20	0.12	46.7	0.42	0.31
BiLDA	40.0	0.25	0.16	38.0	0.48	0.41

Table 3. Intrinsic evaluation of the diversity of the top-5 relevant targets

and for BiDNN are state of the art while BiLDA matching exhibits weaker results. The good baseline results are partly explained by the fact that, in 2015, participants mainly used the textual modality for target selection. Hence the list of targets to rerank contains a significant number of relevant targets with high lexical similarity and using a bag-of-words representation with cosine similarity is adequate. This also explains why baseline results on the top-16 anchors are very strong and, to some extent, why LDA-based approaches fail to be on par with the baseline. Consistently with results in [27], the BiDNN approach performs as well as the baseline, however using a cross-modal approach, showing the interest of autoencoders for multimodal multimedia retrieval. Finally, we note that the post hoc AMT-based annotation process does not encourage diversity. This is beneficial to the baseline, which ranks high segments very similar to the anchor, but detrimental to LDA-based approaches. This consideration also motivates the specific study on diversity described hereunder.

The diversity of the results returned by either one of the methods can be assessed either intrinsically, e.g., by measuring how diverse are the relevant segments found, or by means of subjective human-based judgments. The latter requires that a limited number of anchors be considered to enable a significant number of votes for a single anchor. Diversity is thus evaluated on the 16 anchors for which the best results were obtained with the baseline. For each anchor, we consider the top-5 relevant targets found and test for diversity among these targets. In addition to the lexical and visual representation used for content comparison, we also extracted 10 key words and 10 key concepts for each segment using a tf-idf ranking—a method commonly used as baseline for key word extraction [15]. This compact representation enables comparing the content of the targets and help user quickly apprehend the content of a target.

Table 3 reports a number of intrinsic indicators of the diversity of a list of targets. n_u ($\in [10, 50]$) is the average number of unique key words/concepts in the top-5 relevant segments of an anchor, where the bigger n_u the better the diversity. A value of 10 indicates that all targets have the same key words/concepts. $\overline{d}_a$ is the average cosine similarity between the anchor and the top-5 relevant targets computed over the transcript or over the set of visual concepts. $\overline{d}_i$ measures the similarity within the top-5 targets of an anchor, computed as the average cosine similarity between any two pairs of targets in the top-5 list. In these last two cases, the larger the value, the less diverse the list of 5 targets. Results in Tab. 3 clearly demonstrate that the cross-modal approaches offer a significantly greater diversity of relevant targets than the baseline. Diversity shows both from

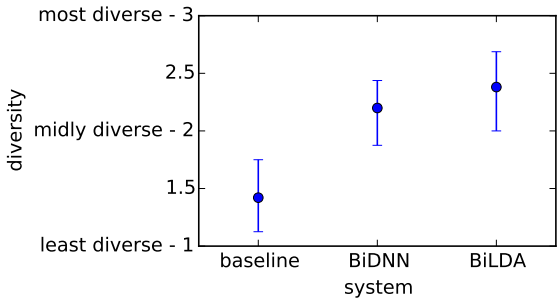

Fig. 4. Average rank of systems with respect to diversity as perceived by evaluators

the lexical standpoint and from the visual one, where the difference between the baseline and cross-modal methods is stronger at the lexical level. BiDNN appears to be slightly better than BiLDA in terms of average distance from targets to anchor as well as in terms of target dispersion.

Results as measured by the intrinsic indicators above are confirmed by user evaluations, where users were presented with an anchor and three lists of 5 relevant targets, one for each method, and asked to rank those lists from the least diverse (rank 1) to the most diverse (rank 3). In the evaluation interface, the anchor appeared on the top of the page, followed by 3 columns of 5 targets each, in a randomized order. Each segment was represented by a key image from which the video could be played, along with 10 key words and 10 key concepts to facilitate the task, potentially avoiding the need to watch all 16 video segments. A session consisted in ranking the lists for the 16 anchors selected, however not all evaluators completed their session. Since the order of anchors was also randomized per session, we kept all votes to report results on as many judgments as possible. In total, 25 persons, mostly from academia, participated in the evaluation, the vast majority of them not familiar with the video hyperlinking task. A total of 176 votes were recorded, with an average of 11 votes per anchor. The annotation took approximately 16 minutes to complete (median time), which corresponds to about one minute per anchor. Results are summarized in Fig. 4 where the average rank is plotted (dot) for each method, with an error bar depicting the dispersion of judgments among users—the lowest/highest average rank assigned to the method by a particular user.

Perceptive evaluations by users confirm the results obtained with intrinsic evaluations, with a significant difference between the transcript-only baseline (average rank of 1.42) and the two cross-modal methods (average ranks of 2.20 and 2.38 for BiDNN and BiLDA resp.). It is also interesting to note that judgments are rather consistent across evaluators, for instance with average ranks from 1.12 to 1.75 for the baseline, confirming the ability of humans to judge diversity. However, contrary to intrinsic evaluations, the relevant targets found by BiLDA were globally perceived as more diverse than those found by BiDNN (significant at $\alpha = 0.01$ according to a paired one-tailed t-test), even though BiLDA performs less than BiDNN in terms of relevance.

5 Conclusion

The study presented in this paper focuses on cross-modal approaches for target selection in video hyperlinking as a mean to offer a diversity of targets. Intrinsic and perceptive evaluations show that cross-modal approaches are significantly better than a text-only baseline at diversity. Bidirectional symmetrical DNNs offer a very good compromise between relevance and diversity. Bimodal LDA offers better potential for diversity but weak performance in terms of relevance still appears as a limitation for this method. However, recent perceptual studies on LDA-derived targets show that combination of topic models can yield performance equivalent to the baseline [23]. Another interesting outcome of the experiments presented here is the fact that diversity can be assessed not only using perceptual tests, but also using intrinsic dispersion measures. The latter are easy to obtain and yield conclusions similar to the one made with the former, opening the door to large-scale studies on diversity in video hyperlinking.

References

1. J. M. Barrios, J. M. Saavedra, F. Ramirez, and D. Contreras. Orand at trecvid 2015: Instance search and video hyperlinking tasks. In *Proc. of TRECVID*, 2015.
2. C. Bhatt, N. Pappas, M. Habibi, and A. Popescu-Belis. Idiap at MediaEval 2013: Search and hyperlinking task. In *Proc. of the MediaEval Workshop*, 2013.
3. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
4. R. Bois, A.-R. Şimon, R. Sicre, G. Gravier, and P. Sébillot. IRISA at TRECVid2015: Leveraging multimodal LDA for video hyperlinking. In *Proc. of TRECVID*, 2015.
5. M. Campr and K. Ježek. Comparing semantic models for evaluating automatic document summarization. In *Text, Speech, and Dialogue*, 2015.
6. Z. Cheng, X. Li, J. Shen, and A. G. Hauptmann. CMU-SMU@TRECVID 2015: Video hyperlinking. In *Proc. of TRECVID*, 2015.
7. T. De Nies, W. De Neve, E. Mannens, and R. Van de Walle. Ghent University-iMinds at MediaEval 2013: an unsupervised named entity-based similarity measure for search and hyperlinking. In *Proc. of the MediaEval Workshop*, 2013.
8. M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The Search and Hyperlinking task at MediaEval 2014. In *Proc. of the MediaEval Workshop*, 2014.
9. M. Eskevich, G. J. Jones, S. Chen, R. Aly, R. Ordelman, D. Nadeem, C. Guinaudeau, G. Gravier, P. Sébillot, T. D. Nies, P. Debevere, R. V. de Walle, P. Galuščáková, P. Pecina, and M. Larson. Multimedia information seeking through search and hyperlinking. In *ACM International Conference on Multimedia Retrieval*, 2013.
10. M. Eskevich, M. Larson, R. Aly, S. Sabetghadam, G. J. F. Jones, R. Ordelman, and B. Huet. Multimodal video-to-video linking: Turning to the crowd for insight and evaluation. In *Proc. of the 23rd International Conference on Multimedia Modeling*, 2017.
11. F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *ACM International Conference on Multimedia*, pages 7–16, 2014.

12. P. Galuscáková, M. Krulis, J. Lokoc, and P. Pecina. CUNI at MediaEval 2014 search and hyperlinking task: visual and prosodic features in hyperlinking. In *Working Notes Proc. of the MediaEval Workshop*, 2014.
13. J.-L. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech communication*, 37(1):89–108, 2002.
14. C. Guinaudeau, G. Gravier, and P. Sébillot. IRISA at MediaEval 2012: Search and hyperlinking task. In *Working Notes Proc. of the MediaEval Workshop*, 2012.
15. K. S. Hasan and V. Ng. Conundrums in unsupervised keyphrase extraction: Making sense of the state-of-the-art. In *Proc. of the 23rd International Conference on Computational Linguistics*, 2010.
16. E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng. Improving word representations via global context and multiple word prototypes. In *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*.
17. H. A. Le, Q. Bui, B. Huet, and et al. LinkedTV at MediaEval 2014 search and hyperlinking task. In *Proc. of the MediaEval Workshop*, 2014.
18. Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proc. International Conference on Machine Learning*.
19. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. of Advances in Neural Information Processing Systems*, 2013.
20. P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, G. Quénot, and R. Ordelman. Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proc. of TRECVID*, 2015.
21. L. Pang and C.-W. Ngo. VIREO @ TRECVID 2015: Video hyperlinking. In *Proc. of TRECVID*, 2015.
22. G. Salton and M. J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, Inc., 1986.
23. A.-R. Simon. *Semantic structuring of video collections from speech: segmentation and hyperlinking*. PhD thesis, Université de Rennes 1, 2015.
24. W. D. Smet and M. Moens. Cross-language linking of news stories on the web using interlingual topic modelling. In *ACM Workshop on Social Web Search and Mining*, 2009.
25. M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440, 2007.
26. T. Tommasi, R. B. N. Aly, K. McGuinness, K. Chatfield, and et al. Beyond metadata: searching your archive based on its audio-visual content. In *Proc. of the International Broadcasting Convention*, 2014.
27. V. Vukotić, C. Raymond, and G. Gravier. Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications. In *Proc. of the ACM International Conference on Multimedia Retrieval*, 2016.
28. V. Vukotic, C. Raymond, and G. Gravier. Multimodal and Crossmodal Representation Learning from Textual and Visual Features with Bidirectional Deep Neural Networks for Video Hyperlinking. In *ACM Multimedia 2016 Workshop: Vision and Language Integration Meets Multimedia Fusion (iV&L-MM'16)*, Amsterdam, Netherlands, Oct. 2016. ACM.
29. I. Vulić and M.-F. Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proc. of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
30. J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.