Zero-Shot Learning of Language Models for Describing Human Actions Based on Semantic Compositionality of Actions

Hideki ASOH
National Institute of
Advanced Industrial Science and Technology
Tsukuba, Ibaraki 305-8568 Japan
h.asoh@aist.go.jp

Ichiro KOBAYASHI
Graduate School of Humanities and Sciences,
Ochanomizu University
Bunkyo-ku, Tokyo 112-8610 Japan
koba@is.ocha.ac.jp

Abstract

We propose a novel framework for zero-shot learning of topic-dependent language models, which enables the learning of language models corresponding to specific topics for which no language data is available. To realize zero-shot learning, we exploit the semantic compositionality of the target topics. Complex topics are normally composed of several elementary semantic components. We found that the language model that corresponds to a particular topic can be approximated with a linear combination of language models corresponding to elementary components of the target topics. On the basis of the findings, we propose simple methods of zero-shot learning. To confirm the effectiveness of the proposed framework, we apply the methods to the problem of generating natural language descriptions of short Kinect videos of simple human actions.

1 Introduction

Constructing topic-dependent language models is useful for many applications such as text mining, speech recognition, statistical machine translation, natural language interfaces, and textual description of images or video contents. In most methods of topic-dependent language model construction, one general model is first constructed from a large amount of language data, and then the general model is modified with a small amount of language data regarding the target topic. The technique of taking the weighted sum of language models is often used for the modification (Bacchiani and Roark, 2003; Jelinek and Mercer, 1980). However, correcting language data for all target topics is demanding and difficult. In particular, when each target topic becomes narrower and the number of target topics increases, it becomes impractical to correct language data for all topics.

In this paper, we propose a novel framework for zero-shot learning of topic-dependent language models, which enables the learning of language models corresponding to specific topics without observing language data regarding the topics on the basis of the semantic compositionality of the target topics.

In the following, we consider rather fine-grained topics such as human activities. Such detailed topics are normally composed of several elementary semantic components. For example, a human action "raising left leg in the forward direction" is considered as a topic. The action includes components such as "up (raise)", "left", "leg", and "in the forward direction". Another action "raising left hand in the side direction" shares the common elements "up" and "left" with the previous action. In this way, actions are related to each other through common components. Hence, the language models generated from natural language sentences describing those actions are also expected to be related to each other. We will show that using this kind of compositionality, we can generate language models corresponding to actions for which we do not have natural language data.

To demonstrate the effectiveness of the proposed methods, we apply the methods to the problem of generating natural language descriptions of short

Kinect videos.

In summary, the original contributions of this work are as follows: 1) the problem of zero-shot learning of topic-dependent language models is newly formulated, 2) novel simple methods for zero-shot learning are proposed, and 3) the effectiveness of the methods is confirmed with real data.

The remainder of the paper is organized as follows: The problem is formalized and solutions are proposed in Section 2, Section 3 discusses related works, Section 4 presents application to the video description problem including experimental setup and results of the experiments, and Section 5 presents the conclusion and discusses future work.

2 Zero-Shot Learning of Language Models

In this section we formalize the problem of zero-shot learning of topic-dependent language models, and propose methods to solve the problem.

2.1 Problem Formalization

As described above, we are interested in the problem of learning multiple topic-dependent language models M_i $(i = 1, ..., N)$, each of which corresponds to a complex fine-grained topic such as human action x_i. When we have a language data S_i i.e. a set of sentences describing the topic x_i for all topics, we can simply calculate M_i from S_i.

The problem we will treat in this paper is estimating language models M_i corresponding to topics x_i for which we do not have language data S_i. Such estimation becomes possible on the basis of the semantic compositionality of topics. We assume that each topic is composed of several semantic components. We denote the semantic components as $y_j (j = 1, ..., K)$.

For example, in the experiments described in Section 4, we use $N = 20$ human actions such as “raising left leg in the forward direction” and “raising both hands in the side direction”. Each action is composed by combining some of $K = 9$ components such as “up”, “down”, “front” (front direction), “side” (side direction), “hand”, “leg”, “right”, “left”, and “both”.

The relation between topics and components can be described by a matrix $A = (a_{ij})$. When $a_{ij} = 1$ then the ith topic includes the jth component, and when $a_{ij} = 0$ then otherwise. In the following section, we assume that a_{ij} is known for all topics. We also assume that the number of topics N is larger than the number of components K.

As for the language model, we consider the n-gram model. An n-gram language model is normally defined by the conditional probabilities $p(w_i|w_{i-1}, ..., w_{i-n+1})$ for a word sequence $(w_{i-n+1}, ..., w_{i-1}, w_i)$. Here we use the joint probabilities $p(w_i, w_{i-1}, ..., w_{i-n+1})$ instead of the conditional probabilities because the joint probabilities are suit for the linear decomposition described below. Hence the conditional probabilities can be calculated from the joint probabilities, this does not reduce the generality and usefulness of the framework.

We denote a vector composed of the joint probability values calculated from language data S_i as ψ_i, and assume that the probability vector ψ_i for the ith topic can be approximately decomposed as the weighted sum of probability vectors ϕ_j corresponding to the jth components included in the topic as

$$\psi_i = \sum_j \frac{a_{ij}}{\sum_j a_{ij}} \phi_j + \varepsilon_i,$$

where ε_i is a vector of the noise term.

Because we consider N topics and K components, the relation can be written with matrices as

$$\Psi = \tilde{A}\Phi + E, \tag{1}$$

where Ψ is an $N \times W$ matrix whose ith row is ψ_i and Φ is a $K \times W$ matrix whose jth row is ϕ_j, and $\tilde{A}$ is a $N \times K$ matrix whose element is $a_{ij} / \sum_j a_{ij}$. W is the dimension of the probability vector of the language model, i.e. the number of ordered word pairs appear in the language data. E is a matrix composed of noise terms. We use this linear relation for zero-shot learning.

2.2 Methods of Zero-Shot Learning

Let us assume we have language data S_i for only N' $(N' < N)$ topics. The set of topics for which we have language data is denoted by T. From the partial language data, we can compute the $N' \times W$ probability vector matrix Ψ' by the same way as the matrix Ψ. A row of Ψ' is the probability vector which corresponds to a topic in T.

If we can estimate Φ for the K components from the partial data, then we can recover the whole Ψ

using the relation of equation (1). This means that we can estimate language models ψ_i for topics for which we have no language data.

We assume that each of K components y_j is included at least once in the N' topics. Then a naive method of computing Φ is to compute the language model ϕ_j from the language data of all topics that include the jth component.

We merge the sentences regarding the topics with the jth component. Then from the merged data we compute the probability vector ϕ_j for the jth component. This method has been designated as "Method 1" in this study.

Another method of estimating Φ is to exploit the least-square estimation to estimate Φ from Ψ' as

$$\hat{\Phi} = \arg\min_{\Phi} ||\Psi' - \tilde{A}'\Phi||^2$$

where $\tilde{A}'$ is an $N' \times K$ matrix made by extracting N' rows corresponding to Ψ' from $\tilde{A}$. This optimization problem can be easily solved as

$$\hat{\Phi} = \tilde{A}'^{+}\Psi',$$

where $\tilde{A}'^{+}$ is the generalized inverse of matrix $\tilde{A}'$. Then from $\hat{\Phi}$ we can estimate the language models for topics without language data. This method has been designated as "Method 2".

3 Related Work

Zero-shot learning has recently become a popular research topic in machine learning, in particular in the domain of large scale visual object recognition and image tagging. Because the number of classes is large, it is difficult to collect true labels for the problems. Hence zero-shot learning is useful in the domain. Palatucci et al. (2009) proposed a method of zero-shot learning and applied to decoding fMRI data from subjects thinking about certain words based on the semantic representation of the target classes. They also gave theoretical analysis of the zero-shot learning framework. Lampert et al. (2009) proposed a method of visual object classification where training and test classes are disjoint. They also exploited semantic attributes of target classes. Farhadi et al. (2009) also proposed rather similar idea.

More recently, Cheng et al. (2013) applied the idea of zero-shot learning to human activity recognition task. They mapped sequence of images to category labels. Socher et al. (2013) proposed a method for zero-shot learning of object recognition using deep neural networks. Frome et al. (2014) improved the model with a larger scale dataset.

All of the previous studies treat zero-shot learning of class labels on the basis of the similarity between input information and also between semantic attribute of the classes. Our work extends the idea of zero-shot learning to language models, which have more complex structure than class labels by exploiting the semantic compositionality of complex topics. In other words, our work goes beyond the word level and treats the sentence level structure. As far as we know, this is the first work which applies the idea of zero-shot learning to topic-dependent language model learning.

The idea of linearly decomposing language models is strongly related to latent topic extraction in text mining. In the latent semantic analysis (LSA), the word frequency vector (unigram probability vector) of a document is linearly decomposed into a weighted sum of latent topic vectors (Deerwester et al., 1990). In topic extraction, the aim of the data analysis is to extract latent topics. On the contrary, in this work, the aim of zero-shot learning is to construct language models for which no language data is available.

In this paper, we assume that the latent topics (= components) are known, and we decompose the language models on the basis of the known combination of components (information of matrix A). However, we can also consider another problem setting where matrix A is unknown. In the setting, the problem is mathematically equivalent with the LSA, and singular value decomposition of the language model matrix Ψ can be used to estimate latent components and language models for the components simultaneously. Various matrix factorization algorithms such as non-negative matrix factorization (Lee and Seung, 1999; Xu et al., 2003), or other probabilistic topic extraction methods such as probabilistic latent semantic analysis (Hofmann, 1999) and latent Dirichlet allocation (Blei et al., 2003) may also be applicable.

Zero-shot learning of language models is also in-

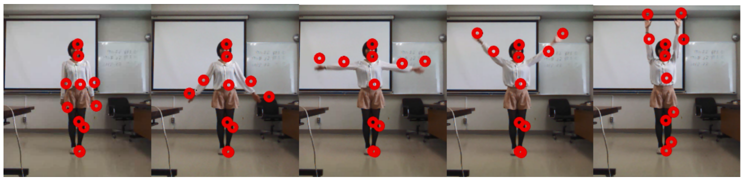

Figure 1: An example of human action (action 11)

teresting from the viewpoint of modeling the natural language acquisition process of humans. Humans are believed to acquire language capability from a rather small amount of observations of language data. To cope with this problem of the poverty of stimuli, certain kinds of zero-shot learning may be exploited. As an example, Sugita and Tani (2005) proposes a model of language acquisition with recurrent neural networks. The robot they constructed can generate sentences describing actions that the robot has not yet experienced on the basis of the semantic compositionality of the actions.

4 Application to Video Content Description System

To demonstrate the effectiveness of the proposed methods, we applied the methods to the problem of generating natural language description of short Kinect videos.

Obtaining a huge amount of video data is becoming easier recently. Whereas we agree with the fact that fully utilization of the data has not been achieved yet. For example, to grasp the content of videos recorded by surveillance cameras, or videos of recorded meetings, we need to watch through the entire videos, which is considerably time-consuming work. If the contents of a video can be recognized and be described with natural language sentences, it will become easier to mine the content of the video data and to achieve various applications such as scene retrieval through natural language queries, etc.

On the basis of such needs, research of the learning relation between natural language and multimedia information has recently been becoming popular in the areas of both natural language processing and multi-media information processing. Many studies have been conducted to generate sentences to explain human behaviors in a video (Barbu et al., 2012; Ding et al., 2012a; Ding et al., 2012b; Kobayashi et al., 2010; Kojima et al., 2002; Rohrbach et al., 2013; Tan et al., 2011). As representative studies, Yu and Siskind (2013) propose a method that learns representations of word meanings from short video clips paired with sentences. Regneri et al. (2013) consider the problem of grounding sentences describing actions in visual information extracted from videos. Takano and Nakamura (2008, 2009) propose incremental learning of association between motion symbols and natural language. Ushiku et al. (2011, 2012) propose a method to create a caption for a still picture, by learning n-gram models for describing picture from pairs of still pictures and their explanation sentences.

Among these works, Kobayashi et al. (2013) are constructing a system for generating natural language description of short Kinect videos of several kinds of human actions. From the pairs of video data of an action taken by the Kinect and Japanese sentences describing the action, the system learns models of observed human actions and language models of the sentences. Using the two models and the correspondence between them, the system can recognizes an action in a new video of a leaned action and outputs Japanese sentences describing the action.

In the work, they assumed that they could collect natural language sentences describing all target actions and construct language models corresponding to all actions from the data. However, when the number of target actions increases, it becomes impractical to prepare natural language descriptions for all actions. Here, we apply our zero-shot learning method to learn the language models of actions for which we do not have language data.

4.1 Experimental Setup

We use $N = 20$ human actions as the target topics. We take short (less than 5 sec.) Kinect videos of

Table 1: Examples of collected sentenses

1	hidari te wo ageru. (raise left hand.)
2	hidari te wo ue ni ageru. (raise left hand upward.)
4	hidari te wo mae kara ageru. (raise left hand to the front direction)
3	hidari te wo shita kara ue ni ageru. (raise left hand upward from below.)
4	hidari te wo mae no hou kara ue ni ageru. (raise left hand upward from the front direction)

the actions, and collect several Japanese sentences that describe the actions. Figure 1 shows an example of an action ("raising both hand through the side direction"). For each action, around 15 sentences describing the action are collected. Table 1 shows some sentences describing the action of raising left hand in the front direction. The collected sentences are segmented into words and bi-gram joint probabilities $p(w_i, w_{i-1})$ are computed from the data for each action. The number of word pairs that appeared in the data is 360.

We set the number of components $K = 9$: i.e., "up", "down", "front" (front direction), "side" (side direction), "hand", "leg", "right", "left", and "both" (only for hands). The combinatorial relationship between the actions and the elements is illustrated in Figure 2. "L", "R", and "B" in the figure denotes "left", "right", and "both" respectively. The figure shows that each human action includes four components in this experiment. For example, Action 3 (ACT 3) is composed of the components "up", "front", "hand", and "left", and Action 18 (ACT 18) is composed of "down", "side", "leg", and "right".

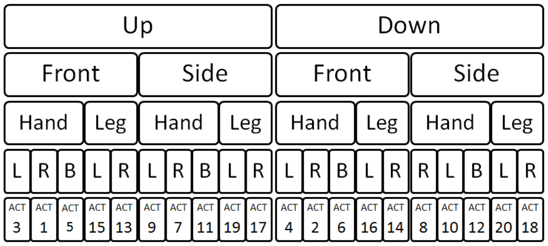

Figure 2: Combinatorial relationship between human actions and components

Table 2: Root mean squared error of the estimated values

Action	Method 1	Method 2	Training	Uniform
1	0.00353	**0.00280**	0.00387	0.00944
2	0.00320	**0.00257**	0.00354	0.00907
3	0.00338	**0.00287**	0.00365	0.00928
4	0.00358	**0.00309**	0.00389	0.00876
5	0.00275	**0.00220**	0.00336	0.00885
6	0.00322	**0.00217**	0.00387	0.00883
7	0.00373	**0.00314**	0.00404	0.00899
8	0.00318	**0.00268**	0.00348	0.00865
9	0.00353	**0.00302**	0.00381	0.00906
10	0.00335	**0.00295**	0.00365	0.00875
11	0.00344	**0.00211**	0.00411	0.00863
12	0.00330	**0.00231**	0.00394	0.00782
13	0.00380	**0.00339**	0.00419	0.00955
14	0.00311	**0.00294**	0.00350	0.00897
15	0.00339	**0.00301**	0.00378	0.00934
16	0.00315	**0.00280**	0.00359	0.00892
17	0.00346	**0.00308**	0.00385	0.00891
18	**0.00297**	0.00301	0.00330	0.00859
19	0.00361	**0.00312**	0.00398	0.00919
20	0.00351	**0.00314**	0.00389	0.00848
Mean	0.00356	**0.00282**	0.00377	0.00890

4.2 Result of Experiment

To evaluate the effectiveness of the proposed zero-shot learning methods, sentences describing one of the 20 human actions are omitted from the training data. Then we estimate Φ for components using $(M-1) \times W$ matrix Ψ' and $(M-1) \times K$ matrix A'. From the estimated $\hat{\Phi}$ we can recover the language model of the sentences omitted from the training data.

Table 1 shows the root mean squared error (RMSE) of the estimated probability values. The column "Action" denotes the target action for which the language data is omitted and the probability vector is estimated with the zero-shot learning methods. The column "Training" means that the language model is estimate using all the sentences in the training data. This is a baseline. Another baseline "Uniform" means that the estimated probability vector is uniform distribution, that is, all probability values are equal to $1/$ (# of word pairs). The minimum RMSE value for each action is shown in bold face.

Compared with the mean value of the non-zero joint probability values 0.0146, it can be said that the

Table 3: Comparisons of the top two most probable sentences

action	With language data	Without language data
1	migi te wo ageru. (raise right hand.) migi te wo ue ni ageru. (move right hand upward.)	migi te wo ageru. (raise right hand.) migi te wo ue ni ageru. (move right hand upward.)
2	migi te wo sageru. (lower right hand.) migi te wo shita ni sageru. (move right hand downward.)	migi te wo sageru. (lower right hand.) migi te wo uekara sageru. (lower right hand from upper position.)
3	hidari te wo ageru. (raise left hand.) hidari te wo ue ni ageru. (move left hand upward.)	hidari te wo ue ni ageru. (move left hand upward.) hidari te wo ageru. (raise left hand.)
5	ryou te wo ageru. (raise both hands.) ryou te wo mae kara ageru. (raise both hands in the forward direction.)	ryou te wo ue ni ageru. (move both hands upward.) ryou te wo ageru (raise both hands.)
18	migi ashi wo orosu. (lower right leg.) migi ashi wo yoko kara orosu. (lower right leg from the side direction.)	migi ashi wo sageru. (lower right leg.) migi ashi wo yoko ni sageru. (lower right leg in the side direction.)
20	hidari ashi wo orosu. (lower left leg.) hidari ashi wo yoko ni orosu. (lower left leg in the side direction.)	hidari ashi wo orosu. (lower left leg.) hidari ashi wo yoko ni orosu. (lower left leg in the side direction.)

RMSE values obtained from our two methods are small enough. The result demonstrates that Method 2 performs better than other methods for allmost all removed topics. However, in Method 2, the estimated values of ϕ_j and ψ_i do not become probabilities, that is, some values may become below zero and the sum of the values slightly differ from one. Hence, it becomes a bit difficult to interpret the values. Although this is not so serious problem in practice, this can be considered as a kind of tradeoff between the accuracy and the interpretability.

We also evaluate the RMSE values when we omit language data for more than one actions from the training data. The results strongly depend on the data which are omitted. For example, when we omit language data regarding actions 1, 2, 7, and 8, then the RMSE value of the estimated language model for Action 1 is degraded to 0.00469. However when we omit language data regarding actions 1, 3, 5, and 13, then the RMSE keeps low value 0.00223.

This difference comes from the components included in the remaining actions. The Action 1 is composed of "raise", "front", "right", "hand". When we omitted actions 1, 2, 7, and 8, no actions including components "right" and "hand" is remained in the training data. Hence this causes rather serious effect to the accuracy of the zero-shot estimation. However, when we omitted actions 1, 3, 5, and 13, all component pairs are still included in the training data. Hence this does not cause serious damage to the estimated language model.

Through the analysis of various cases, we confirmed that if the choice of omitted data is balanced to keep all semantic components remained in the training data, then the performance of zero-shot learning is not degraded so much even though language data regarding several actions are omitted.

Finally we evaluate the text generation capability of the estimated language models. Here we use the language models estimated by Method 2. We generate Japanese sentences of high likelihood value in the same way as in the work of Kobayashi et al. (2013), i.e. with the Viterbi algorithm using the language model of each action.

Table 3 contrasts the top two most probable texts generated with the bi-gram computed from the collected language data of the action and with the bi-gram estimated by the zero-shot learning using the language data of the other 19 actions. We demonstrate the results for 6 of the 20 actions. From the table, we can see that almost the same sentences are generated with the bi-gram probability vector estimated by our zero-shot learning method.

Although the actions used in the experiment are rather simple, we confirmed the possibility of zero-shot learning of effective language models. Those results show that zero-shot learning is a promising way to cope with the problem of the poverty of language data in natural language processing.

5 Conclusion and Future Work

We have proposed methods of zero-shot learning of fine-grained topic-dependent language models. Using the methods, we can learn topic-dependent language models corresponding to topics for which we do not have language data on the basis of the compositionality of the topics. We confirmed the effectiveness of the proposed methods with the task of describing short Kinect videos of human actions.

Much work remains to be done in the future. Because our experiment was conducted with a small-scale dataset, the methods should be evaluated more elaborately with larger scale datasets. The proposed zero-shot learning may be useful not only for describing videos but also for other various applications such as speech recognition, machine translation, text mining, and video retrieval. Application of the methods to such problems is an interesting topic.

In this paper, we assumed that the matrix A which denotes the relationship between actions and components is known. However, as is mentioned in the related work section, the problem setting for unknown A is also interesting. This problem is related to find the optimal elementary components to describe target topics. This is a kind of dictionary learning problem.

Finally, modeling more complex relation between multiple language models using more sophisticated probabilistic models may be an interesting research direction for natural language processing. As an example, Eisenstein et al. (2011) proposed a new way of representing multiple language models. Introducing their method of sparse additive decomposition of language models into our framework is also an interesting issue.

Acknowledgement: This work is partly supported by JSPS KAKENHI 26280096, MEXT KAKENHI 25120011, and foundation for the Fusion of Science and Technology.

References

M. Bacchiani and B. Roark. 2003. Unsupervised Language Model Adaptation. *2003 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol.1:224–227.

A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang. 2012. Video In Sentences Out. *arXiv:1204.2742*.

D. M. Blei, A. Y. Ng, and M. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3 (4–5): 993–1022.

H.-T. Chang, M. Griss, P. Davis, J. Li, and D. You. 2013. Towards Zero-Shot Learning for Human Activity Recognition Using Semantic Attribute Sequence Model. *Proceedinsg of UbiComp'13*.

A. Farhadi, I. Endres, D. Holem, and D. Forsyth. 2009. Describing Objects by Their Attributes. *Proceedings of CVPR 2009*.

A. F. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. 2014. DeViSE: A Deep Visual-Semantic Embedding Model, *Proceedings of NIPS 2014*.

S. Deerwester, S. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41 (6): 391–407.

D. Ding, F. Metze, S. Rawat, P. F. Schulam, and S. Burger. 2012. Generating Natural Language Summaries for Multimedia. *Proceedings of the 7th International Natural Language Generation Conference*, 128–130.

D. Ding, F. Metze, S. Rawat, P. F. Schulam, S. Burger, E. Younessian, L. Bao, M. G. Christel, and A. Hauptmann. 2012. Beyond Audio and Video Retrieval: Towards Multimedia Summarization. *Proceeding of the 2nd ACM International Conference on Multimedia Retrieval*, Article No.2.

J. Eisenstein, A. Ahmed, and E. P. Xing. 2011. Sparse Additive Generative Models of Text. *Proceedings of the 28th International Conference on Machine Learning*.

T. Hofmann. 1999. Probabilistic Latent Semantic Analysis. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 289-296.

F. Jelinek and R. L. Mercer. 1980. Interpolated Estimation of Markov Source Parameters from Sparse Data. *Proceedings of the Workshop on Pattern Recognition in Practice*.

I. Kobayashi, M. Noumi, and A. Hiyama. 2010. A Study on Verbalization of Human Behaviors in a Room. *Proceedings of the 2010 IEEE International Conference on Fuzzy Systems*.

M. Kobayashi, I. Kobayash, H. Asoh, and S. Guadarrama. 2013. A Probabilistic Approach to Text Generation of Human Motions Extracted from Kinect Videos. *Proceedings of the World Congress on Engineering and Computer Science 2013*.

A. Kojima, T. Tamura, and K. Fukunaga. 2002. Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions. *International Journal of Computer Vision*, 50 (2):171–184.

C. H. Lambert, H. Nickisch, and S. Harmeling. 2009. Learning to Detect Unseen Object Classes by Between-Class Attribute Transfer. *Proceedings of CVPR 2009*.

D. D. Lee and H. S. Seung. 1999. Learning the Parts of Objects with Nonnegative Matrix Factorization. *Nature*, 401, 788–791.

M. Palatucci, D. Pomerleau, G. Hinton, and T. M. Mitchell. 2009. Zero-shot Learning with Semantic Output Codes. *Proceedings of NIPS 2009*.

M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. 2013. Grounding Action Descriptions in Videos. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. 2013. Translating Video Content to Natural Language Descriptions. *Proceedings of ICCV 2013*, 433-440.

R. Socher, M. Ganjoo, C. D, Manning, and A. Y. Ng. 2013. Zero-shot learning through cross-modal transfer. *Proceedings of NIPS 2013*.

Y. Sugita and J. Tani. 2005. Learning Semantic Combinatoriality from the Interaction between Linguistic and Behavioral Processes. *Adaptive Behaviour*, 3 (1): 33–52.

W. Takano and Y. Nakamura. 2008. Integrating Whole Body Action Primitives and Natural Language for Humanoid Robots. *Proceedings of 2008 IEEE-RAS International Conference on Humanoid Robots*, 708–713.

W. Takano and Y. Nakamura. 2009. Incremental Learning of Integrated Semiotics based on Linguistic and Behavioral Symbols. *Proceedings of 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1780–1785.

C. C. Tan, Y.-G. Jiang and C.-W. Ngo. 2011. Towards Textually Describing Complex Video Contents with Audio-Visual Concept Classifiers. *Proceedings of the 19th ACM international conference on Multimedia*, 655–658.

Y. Ushiku, T. Harada, and Y. Kuniyoshi. 2011. A Understanding Images with Natural Sentences. *Proceedings of the 19th Annual ACM International Conference on Multimedia*, 679–682.

Y. Ushiku, T. Harada, and Y. Kuniyoshi. 2012. Efficient Image Annotation for Automatic Sentence Generation. *Proceedings of the 20th Annual ACM International Conference on Multimedia*, 549–558.

W. Xu, X. Liu, and Y. Gong. 2003. Document Clustering based on Non-negative Matrix Factorization. *Proceedings of 26th Annual International ACM SIGIR Conference*, 267–273.

H. Yu, and J. M. Siskind. 2013. Grounded Language Learning from Video Described with Sentences. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.