# Learning Word Meanings and Grammar for Describing Everyday Activities in Smart Environments

**Muhammad Attamimi**[1] **Yuji Ando**[1] **Tomoaki Nakamura**[1] **Takayuki Nagai**[1]
**Daichi Mochihashi**[2] **Ichiro Kobayashi**[3] **Hideki Asoh**[4]
[1] The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo, Japan
[2] Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo, Japan
[3] Ochanomizu University, 2-1-1 Otsuka Bunkyo-ku Tokyo, Japan
[4] National Institute of Advanced Industrial Science and Technology,
1-1-1 Umezono, Tsukuba, Ibaraki, Japan
{m_att, ando, naka_t}@apple.ee.uec.ac.jp, tnagai@ee.uec.ac.jp,
daichi@ism.ac.jp, koba@is.ocha.ac.jp, h.asoh@aist.go.jp

## Abstract

If intelligent systems are to interact with humans in a natural manner, the ability to describe daily life activities is important. To achieve this, sensing human activities by capturing multimodal information is necessary. In this study, we consider a smart environment for sensing activities with respect to realistic scenarios. We next propose a sentence generation system from observed multimodal information in a bottom up manner using multilayered multimodal latent Dirichlet allocation and Bayesian hidden Markov models. We evaluate the grammar learning and sentence generation as a complete process within a realistic setting. The experimental result reveals the effectiveness of the proposed method.

## 1 Introduction

Describing daily life activities is an important ability of intelligent systems. In fact, we can use this ability to achieve a monitoring system that is able to report on an observed situation, or create an automatic diary of a user. Recently, several studies have been performed to generate sentences that describe images using Deep Learning (Vinyals et al., 2014; Fang et al., 2014; Donahue et al., 2014; Kiros et al., 2015). Although these results were good, we are interested in unsupervised frameworks. This is necessary to achieve a system that can adapt to the user, that is, one that can learn a user-unique language and generate it automatically. Moreover, the use of crowdsourcing should be avoided to respect the privacy of the user. Regarding this, studies on sentence generation from RGB videos have been discussed in (Yu and Siskind, 2013; Regneri et al., 2013). A promising result for language learning has been shown in (Yu and Siskind, 2013) and a quite challenging effort to describe cooking activities was made in (Regneri et al., 2013). However, these studies rely only on visual information, while we aim to build a system that is able to describe everyday activities using multimodal information. To realize such systems, we need to consider two problems. The first problem is the sensing of daily life activities. In this paper, we utilize a smart house (Motooka et al., 2010) for sensing human activities. Thanks to the smart house, multimodal information such as visual, motion, and audio data can be captured. The second problem to be tackled is verbalization of the observed scenes. To solve this problem, a multilayered multimodal latent Dirichlet allocation (mMLDA) was proposed in (Attamimi et al., 2014).

In this paper, we propose a sentence generation system from observed scenes in a bottom up manner using mMLDA and a Bayesian hidden Markov model (BHMM) (Goldwater and Griffiths, 2007). To generate sentences from scenes, we need to consider the words that represent the scenes and their order. Here, mMLDA is used to infer words for given scenes. To determine the order of words, inspired by (Kawai et al., 2014), a probabilistic grammar that considers syntactic information is learned using BHMM. In this study, the order of concepts is generated by sampling the learned grammar. The word selection for each generated concept is then performed using the observed data. Moreover, a language model that represents the relationship between words is also used to calculate



*Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2254,
Lisbon, Portugal, 17-21 September 2015. 

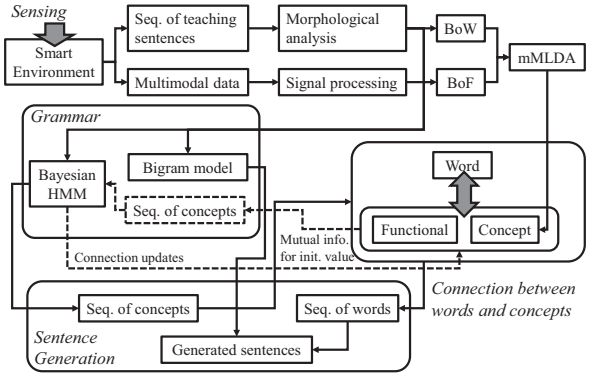


Figure 1: Language learning and sentence generation system.

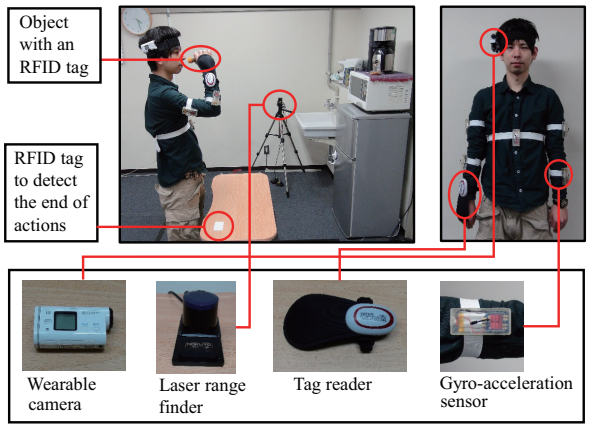


Figure 2: Multimodal information acquisition.

the transition probability between them. Considering the transition probability at word level, a lattice of word candidates corresponding to the concept sequence can be generated. Therefore, sentence generation can be thought of as a problem of finding the word sequence that has the highest probability from the lattice of word candidates, which can be solved by the Viterbi algorithm. Finally, sampling from grammar is performed multiple times to generate sentence candidates and select the most probable one.

## 2 Proposed method

### 2.1 Overview

Figure 1 illustrates the overall system of proposed language learning and sentence generation. In this study, we use a smart environment for sensing multimodal information. The system shown in Figure 2 is part of a smart house (Motooka et al., 2010) that is used to capture multimodal information. Here, an RFID tag is attached to an object

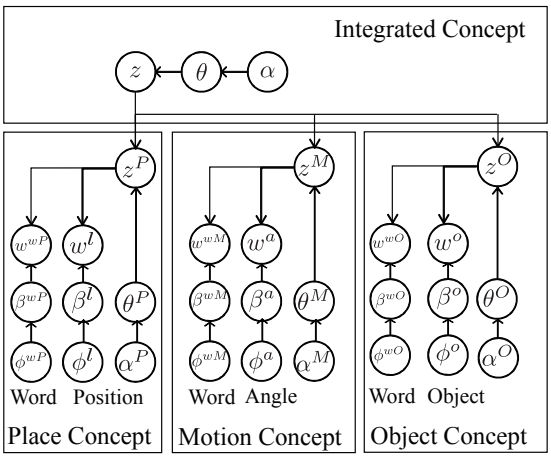


Figure 3: Graphical model of mMLDA.

to enable the object information to be read using a wearable tag reader. To capture motion, five sensors that consist of 3-axis acceleration with 3-axis gyroscope sensors are attached to the upper body, as shown in Figure 2. Moreover, a particle filter-based human tracker (Glas et al., 2007) applied to four laser range finders is used to estimate the location of a person while performing an action. This is a setup designed to demonstrate that language can be learned and generated from real human actions. Ultimately, our goal is sensing based on image recognition.

The acquired multimodal data is then processed, which results in a bag-of-words model (BoW) and bag-of-features model (BoF) (Csurka et al., 2004). Using mMLDA (see section 2.2), various concepts can be formed from the multimodal data. Given teaching sentences, the connection between words and concepts can be learned based on mMLDA and BHMM which is learned with mutual information (MI) as the initial value. On the other hand, the bigram model of words is calculated and used as the score when reordering words inferred from multimodal information using grammar. A morphological analyzer for parsing words in a sentence is also necessary in the proposed system. We use publicly available parser MeCab (Kudo et al., 2004). In the future, we plan to use the unsupervised morphological analysis technique proposed in (Mochihashi et al., 2009).

### 2.2 mMLDA

Figure 3 shows the graphical model of mMLDA used in this paper. Here, $z$ represents the integrated category (concept), whereas $z^O$, $z^M$, and $z^P$ represent the object, mo-

tion, and place concepts, respectively. In the bottom layer (lower panel of Figure 3), $w^m \in \{w^o, w^{wO}, w^a, w^{wM}, w^l, w^{wP}\}$ represents the multimodal information obtained from each object, motion, and place. Here, $w^o$, $w^a$, and $w^l$ denote multimodal information obtained respectively from the object used in an action, motion of a person while using the object, and location of the action. Further, $w^{wC} \in \{w^{wO}, w^{wM}, w^{wP}\}$ denotes word information obtained from teaching sentences. Observation information is acquired by using the system shown in Figure 2. A brief explanation of each observation is as follows.

For object information, an $N_o$-dimensional vector $w^o = (o_1, o_2, \cdots, o_{N_o})$ is used, where $N_o$ denotes the number of objects. In this vector, $o_*$ takes a value of 0 or 1, where $o_i$ is set to 1 if an object with index $i$ is observed. Moreover, all of the teaching sentences are segmented into words and represented by a BoW as word information. Here, motion is segmented according to the object used. A sequence of 15-dimensional feature vectors for each motion is acquired. Using BoF, the acquired feature vectors are vector quantized, resulting in a 70-dimensional vector. The acquired two dimensional of human positions are processed using BoF to construct a 10-dimensional vector as place information.

In mMLDA, latent variables that represent upper and lower concepts $z$ and $z^C \in \{z^O, z^M, z^P\}$ are learned simultaneously. Gibbs sampling is applied to the marginalized posterior probability of latent variables to learn the model from observed data $w^m$ (Attamimi et al., 2014).

## 2.3 Language learning and generation

### 2.3.1 Word inference

In this study, word information is obtained from teaching sentences and employed for all concepts, as shown in Figure 3. Considering that appropriate words to express each concept exist, a criterion to measure the correlation between words and concepts is needed. At the start of grammar learning, MI, which can measure the mutual dependence of two stochastic variables, is used. Therefore, a word is considered to express a category when the MI between the word and category is large. On the other hand, a word with small MI is identified as a functional word. This determination is used as an initial value in the syntactic learning and needs not be strictly determined. Once the grammar is learned, we can utilize BHMM's parameters $P(w^w|c)$ to infer a word $w^w$ from observed data $\boldsymbol{w}^m_{\text{obs}}$ as $\hat{P}(w^{wC}|\boldsymbol{w}^m_{\text{obs}}, c) \propto \max_k P(w^{wC}|c)P(w^{wC}|k)P(k|\boldsymbol{w}^m_{\text{obs}}, c)$, where $P(w^{wC}|k)$ and $P(k|\boldsymbol{w}^m_{\text{obs}}, c)$ can be estimated from mMLDA (Attamimi et al., 2014) and $k$ is category of concept $c' \in \{object, motion, place\}$ and $c \in \{c', functional\}$. It should be note that $P(w^{wC}|k)$ and $P(k|\boldsymbol{w}^m_{\text{obs}}, c)$ are considered as uniform distribution for "functional" since they cannot be inferred from observed data using mMLDA. In this case, we can rely on syntactic information which is learned by BHMM.

### 2.3.2 Grammar learning using BHMM

Thanks to mMLDA and BHMM, appropriate words to represent the observed information can be inferred. Given an input consisting of a teaching sentence of a sequence of words, a BHMM can infer a sequence of concepts. In the learning phase, the MI results of concept selection for each word are used as the initial values of the BHMM. Here, grammar is defined as the concept transition probability $P(C_t|C_{t-1})$, which is estimated using Gibbs sampling, where $C_t \in c$ represents the corresponding concepts of the $t$-th word in the sentence. In addition, a language model that represent the bigram model of words in the teaching sentences is also used for generating sentences.

| Motion | Object | Place | Motion | Object | Place | Motion | Object | Place |
|---|---|---|---|---|---|---|---|---|
| Drink (1) | Juice (1) | Sofa (1) | Wipe (7) | Dustcloth (9) | Kitchen (4) | Write on (12) | Notebook (16) | Bedroom (5) |
| | Tea (4) | Dining room (2) | | Tissue (10) | Dining room (2) | | Textbook (17) | Sofa (1) |
| Eat (2) | Cookies (2) | Dining room (2) | Turn on (8) | Remote control | Living room (3) | | Refrigerator (18) | Kitchen (4) |
| | Chocolate (3) | Living room (3) | | (air conditioner) (11) | Bedroom (5) | Open (13) | Microwave (19) | Kitchen (4) |
| Shake (3) | Tea (4) | Sofa (1) | Open (turn) (9) | Tea (4) | Living room (3) | | Closet (20) | Bedroom (5) |
| | Dressing (5) | Kitchen (4) | | Honey (6) | Dining room (2) | Read (14) | Textbook (17) | Bedroom (5) |
| Pour (4) | Tea (4) | Kitchen (4) | Wrap (10) | Plastic wrap (12) | Dining room (2) | | Magazine (21) | Sofa (1) |
| | Juice (1) | Living room (3) | | Aluminum foil (13) | Kitchen (4) | Spray (15) | Deodorizer (22) | Living room (3) |
| Put on (5) | Dressing (5) | Dining room (2) | | Shirt (14) | Bedroom (5) | | | Bedroom (5) |
| | Honey (6) | Kitchen (4) | Hang (11) | | | | Scourer (23) | Kitchen (4) |
| Throw (6) | Ball (7) | Sofa (1) | | Parka (15) | Living room (3) | Scrub (16) | Sponge (24) | Kitchen (4) |
| | Plushie (8) | Bedroom (5) | | | | | | |

Table 1: Object, motion, and place correspondences (numbers in parentheses represent the category index).

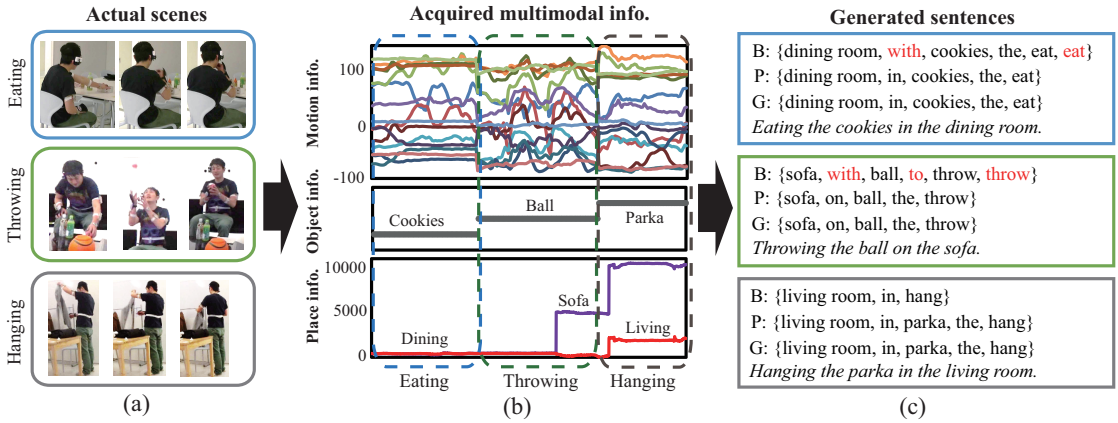


Figure 4: Examples of: (a) actual images, (b) captured multimodal information, and (c) generated sentences. In each image, B, P, and G indicate the sentence structure in Japanese grammar generated by the baseline method, proposed method, and correct sentence, respectively; whereas the bottom line gives the meaning of the generated sentence. Words marked in red have been incorrectly generated.

#### 2.3.3 Sentence generation of observed scenes

First, concepts are sampled from the begin of sentence "BOS" until the end of sentence "EOS" according to the learned grammar $N$ times. Let the $n$-th ($n \in \{1, 2, \cdots, N\}$) sequence of concepts that excludes "BOS" and "EOS" be $\boldsymbol{C}^n = \{C_1^n, \cdots, C_t^n, \cdots, C_{T_n}^n\}$, where, $T_n$ denotes the number of sampled concepts, which corresponds to the length of a sampled sentence.

Next, the word that corresponds to concept $C_t^n$ is estimated. Here, for a given observed information $\boldsymbol{w}_{\text{obs}}^m$, the top-$K$ words that correspond to concept $C_t^n$ and have high probabilities $\boldsymbol{w}_t^n = \{w_{t1}^n, w_{t2}^n, \cdots, w_{tK}^n\}$ are used. Hence, the set of all words for a sequence of concepts $\boldsymbol{C}^n$ can be written as $\boldsymbol{W}^n = \{\boldsymbol{w}_1^n, \boldsymbol{w}_2^n, \cdots, \boldsymbol{w}_{T_n}^n\}$. Therefore, $K^{T_n}$ number of patterns for a candidate of the sentence can be considered for $\boldsymbol{C}^n$ and the corresponding words $\boldsymbol{W}^n$. Each candidate for sentence $S^n$ is selected from these patterns and has the following probability:

$$P(S^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}_{obs}^m) \propto \prod_t P(C_t^n|C_{t-1}^n) P(w_t^n|\boldsymbol{w}_{obs}^m, C_t^n) P(w_t^n|w_{t-1}^n). \quad (1)$$

For observed information, the most probable sentence is selected from $N$ sequences of concepts with sets of words. Here, the sentence $\hat{S}^n$ that maximizes Eq. (1) is determined for each sequence of concepts. Because many patterns of $S^n$ exist, the Viterbi algorithm is applied to cut the computational cost and determine the most probable sentence. Thus, a set of sentences that consists of sentences with the highest probability for each sequence of concepts can be written as $\hat{\boldsymbol{S}} = \{\hat{S}^1, \cdots, \hat{S}^n, \cdots, \hat{S}^N\}$.

We can select the final sentence from $\hat{\boldsymbol{S}}$ by considering the most probable candidate. In fact, long sentences tend to have low probability and are less likely to be selected. To cope with this problem, adjustment coefficient $\ell(\hat{S}^n) = \frac{(L^{\max} - L_{\hat{S}^n})}{\sum_n^N L_{\hat{S}^n}} \sum_n^N \log P(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}_{obs}^m)$ is introduced, where, $L_{\hat{S}^n}$ denotes the length of sentence $\hat{S}^n$ and $L^{\max}$ represents the maximum value of the sentence length in $\hat{\boldsymbol{S}}$. Using $\ell(\hat{S}^n)$, the logarithmic probability of the sentence can be calculated as $\log \bar{P}(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}_{obs}^m) = \log P(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}_{obs}^m) + \omega \ell(\hat{S}^n)$, where $\omega$ is a weight that controls the length of sentences. A large weight leads to longer sentences. The final sentence $S$ is determined as $S = \text{argmax}_{\hat{S}^n \in \hat{\boldsymbol{S}}} \log \bar{P}(\hat{S}^n|\boldsymbol{C}^n, \boldsymbol{W}^n, \boldsymbol{w}_{obs}^m)$.

## 3 Experiments

The acquisition system shown in Figure 2 was used to capture multimodal information from human actions. Table 1 shows the actions that were performed by three subjects twice, resulting in a total of 195 multimodal data with 1170 sentences. We then divided the data into training data (99 multimodal data with 594 sentences) and test data (96 multimodal data with 576 sentences). Some examples of acquired multimodal data are shown in Figure 4(b). Using training data, various concepts were formed by mMLDA, and the categorization accuracies for object, motion, and place were respectively 100.00%, 52.53%, and 95.96%. Motion similarity was responsible for the false cat-

| | ♯ of words | Baseline | Proposed |
|---|---|---|---|
| w/o functional words | 78 | 65.38% | **73.08%** |
| w functional words | 98 | – | **68.37%** |

Table 2: Concepts selection results.

egorization of motion concepts. Since our goal is to generate sentences from observed scenes, these results are used as reference instead of comparing with the baseline.

To evaluate the concept selection of words, 98 words in teaching sentences were used. We compared the results of concept selection with hand-labeled ones. Table 2 shows the accuracy rate of concept selection. Here, we excluded the functional words (resulting in 78 words) for fair comparison with the baseline method (Attamimi et al., 2014). One can see that, better results can be achieved by the proposed method. It is clear that concept selection is improved by using the BHMM, indicating that a better grammar can be learned using this model.

Next, the learned grammar was used and sentences were generated. To reduce randomness of the results, sentence generation was conducted 10 times for each data. To verify sentence generation quantitatively, we evaluated the sentences automatically using BLEU score (Papineni et al., 2002). Figure 5 depicts the results of 2- to 4-gram of BLEU scores. Since functional words are not considered in (Attamimi et al., 2014), we used our grammar and performed sentence generation proposed in (Attamimi et al., 2014) as the baseline method. One can see from the figure that in all cases the BLEU scores of proposed method out performs the baseline method. It can be said that the sentences generated by the proposed method are of better quality than those generated by the baseline method.

Moreover, we also manually evaluated generated sentences by asking four subjects (i.e., college students who understand Japanese) whether the sentences were: correct both in grammar and meaning (E1), grammatically correct but incorrect in meaning (E2), grammatically incorrect but correct in meaning (E3), or incorrect both in grammar and meaning (E4). The average rates of E1, E2, E3, and E4 were shown in Table 3. We can see that the proposed method out performs the baseline method by providing high rates of E1 and E2; and low rates of E4. Because we want to generate sentences that explain actions, incorrect motion inference would lead to incorrect sentence generation. Examples of E2 are "Eating the plastic wrap in the dining room" and "Opening the dressing in the kitchen." One can see that these sentences are grammatically correct but do not express the scenes correctly because the words that represent the motion are incorrect. Hence, the misclassification that occurred in the motion concept formation was responsible for the incorrect meaning of the generated sentences. Figure 4(c) shows the sentences generated from the given scenes (Figure 4(a)). We can see that meaningful yet natural sentences that explain the observed scenes can be generated using the proposed method.

| | Grammar | Meaning | Baseline | Proposed |
|---|---|---|---|---|
| E1 | **correct** | **correct** | $(23.21 \pm 5.28)\%$ | $(45.39 \pm 3.02)$**%** |
| E2 | **correct** | incorrect | $(35.07 \pm 9.32)\%$ | $(49.79 \pm 3.77)$**%** |
| E3 | incorrect | **correct** | $(11.34 \pm 5.59)$**%** | $(2.79 \pm 2.39)\%$ |
| E4 | incorrect | incorrect | $(30.38 \pm 10.54)$**%** | $(2.03 \pm 2.10)\%$ |

Table 3: Evaluation results of generated sentences.

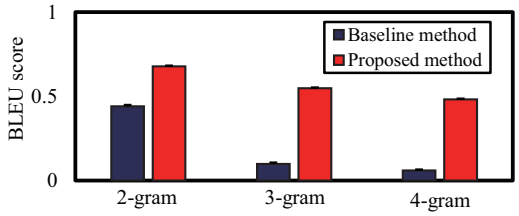


Figure 5: BLEU scores of generated sentences.

## 4 Conclusion

In this paper, we proposed an unsupervised method to generate natural sentences from observed scenes in a smart environment using mMLDA and BHMM. In the smart environment, multimodal information can be acquired for realistic scenarios. Thanks to mMLDA, various concepts can be formed and an initial determination of functional words can be made by assuming a weak connection of concepts and words calculated by MI. The possibility that grammar can be learned from BHMM by considering the syntactic information has also been shown. We conducted experiments to verify the proposed sentence generation, and promising preliminary results were obtained. In future work, we aim to implement a nonparametric Bayes model that will be able to estimate the number of concepts automatically.

## Acknowledgments

This work is partly supported by JSPS KAKENHI 26280096.

## References


Muhammad Attamimi, Muhammad Fadlil, Kasumi Abe, Tomoaki Nakamura, Kotaro Funakoshi, and Takayuki Nagai. 2014. *Integration of Various Concepts and Grounding of Word Meanings Using Multi-layered Multimodal LDA for Sentence Generation*. In Proc. of IEEE/RSJ International Conference on Intelligent Robots, pp.2194–2201.

Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, Cédric Bray. 2004. *Visual Categorization with Bags of Keypoints*. In Proc. of ECCV International Workshop on Statistical Learning in Computer Vision.

Jeffrey Donahue, Lisa Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2014. *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*. Technical Report No. UCB/EECS-2014-180.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2014. *From Captions to Visual Concepts and Back*. In Proc. of IEEE International Conference on Computer Vision and Pattern Recognition.

Dylan F. Glas, Takahiro Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2007. *Laser Tracking of Human Body Motion Using Adaptive Shape Modeling*. In Proc. of IEEE/RSJ International Conference on Intelligent Robots, pp.602–608.

Sharon Goldwater and Thomas L. Griffiths. 2007. *A Fully Bayesian Approach to Unsupervised Part-of-Speech Tagging*. In Proc. of the 45th Annual Meeting of the Association for Computational Linguistics, pp.744–751.

Yuji Kawai, Yuji Oshima, Yuki Sasamoto, Yukie Nagai, and Minoru Asada. 2014. *Computational Model for Syntactic Development: Identifying How Children Learn to Generalize Nouns and Verbs for Different Languages* In Proc. of Joint IEEE International Conferences on Development and Learning and Epigenetic Robotics, pp.78–84.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2015. *Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models*. In Trans. of the Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. *Appliying Conditional Random Fields to Japanese Morphological Analysis*. In Proc. of Conference on Empirical Methods in Natural Language Processing, pp.230–237.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. *Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling* In Proc. of the Association for Computational Linguistics, pp.100–108.

Nobuhisa Motooka, Ichiro Shiio, Yuji Ohta, Koji Tsukada, Keisuke Kambara, and Masato Iguchi. 2010. *Ubiquitous Computing House Project: Design for Everyday Life*. *Journal of Asian Architecture and Building Engineering*, 8:77–82.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. In Proc. of the Association for Computational Linguistics, pp.311–318.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. *Grounding Action Descriptions in Videos*. Trans. of the Association for Computational Linguistics, 1:25–36.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. *Show and Tell: A Neural Image Caption Generator*. In arXiv:1411.4555 [cs.CV].

Haonan Yu and Jeffrey M. Siskind. 2013. *Grounded Language Learning from Video Described with Sentences*. In Proc. of the Association for Computational Linguistics, pp.53–63.