MMG: A Learning Game Platform for Understanding and Predicting Human Recall Memory

Umer Fareed[1] and Byoung-Tak Zhang[1,2]

[1] School of Computer Science and Engineering
[2] Graduate Programs in Cognitive Science and Brain Science
Seoul National University, Gwanak-gu, Seoul 151-742, Korea
{ufareed,btzhang}@bi.snu.ac.kr

Abstract. How humans infer probable information from the limited observed data? How they are able to build on little knowledge about the context in hand? Is the human memory repeatedly constructing and reconstructing the events that are being recalled? These are a few questions that we are interested in answering with our multimodal memory game (MMG) platform that studies human memory and their behaviors while watching and remembering TV dramas for a better recall. Based on the preliminary results of human learning obtained from the MMG games, we attempt to show that the human memory recall improves steadily with the number of game sessions. As an example case, we provide a comparison for the text-to-text and text-image-to-text learning and demonstrate that the addition of image context is useful in improving the learning.

Keywords: human learning, memory recall, Bayesian inference, human cognition.

1 Introduction

With the significant advancement in the field of artificial intelligence and machine learning over the past two decades, people tend to consider machines to be the near future replacement for many human tasks [1, 2]. But still, one can think of many tasks that humans perform better than the machines. The human ability to make robust, flexible and reliable inferences from limited data is one of the indisputable cases where nobody is able to find any counter argument. Without any doubt, there is a vast gap between human and machine learning at many levels. Take the example of language acquisition; being one of the hardest tasks for the computing machines to decipher but humans succeed in learning language purely from linguistic input. What we need here is to develop and improve computational systems for bridging this gap between humans and machines. These developments surely can contribute towards clearer and deeper insight into human cognition principles.

For instance, when we consider the case wherein people recognize words so quickly and so accurately from noisy speech, we ought to investigate what helps them to parse a string of words from the underlying grammar. Most of the time, inadequate data severely limits the inferences which people make about the data. Many of these

B.-H. Kang and D. Richards (Eds.): PKAW 2010, LNAI 6232, pp. 300–309, 2010.

choices come from prior probabilities of word structures known to them in one context or another. This thought drives us to conclude that the nature of the constraints on human inferences that guide human learning could be formalized using probabilistic models of cognition [3, 4].

Recently, there has been much discussion about probabilistic models of cognition. These techniques include probabilistic graphical models, such as Bayesian networks [5], Markov random fields [6], and Markov logic networks [7]. More recent models include deep belief networks [8] and random hypergraph structures [9] for parallel associative memory.

In our proposal, we describe a multimodal memory game (MMG) platform that studies human learning and memory behaviors to play and the machine to learn in an interactive digital cinema environment. The game manager being the core component of the game is responsible for controlling the working and learning environment of the game. We try to focus on machine learning with the help of human players whom we give a few minutes TV program clip to watch. Later, we ask questions about the subsequent text for the given text or text-image pairs taken from the video corpus. The human player's answers to the text or text-image queries teach the machine in the learning process to generate the succeeding text for different learning scenarios. We plot the points scored by the human players against the number of sessions played by them and observe the learning accuracy (performance of accurate recall) for the human players. Our interest is in modeling this learning accuracy and investigating how this learning leads to a normalized behaviour for learned vision-language modality information.

The proposed framework of the multimodal game is flexible and can be adapted to a more complex situation where more modalities and additional users are incorporated. The game also provides a research platform to study the life-long learning process in a dynamic environment since the video data can be played in a varying scenarios. The use of multimodal memory game for the experiments is a step towards the realization that cognitive learning associated with the human memory can be better understood with practical implementation through machine learning processes [10, 11, 12].

2 Human Learning and the Bayesian Models

Human learning in the real world is inductive, i.e., the learner builds generalizations or makes predictions based on the limited information in hand. Moreover, latest computational models of visual perception and inductive inference have demonstrated that Bayesian framework can be the optimal choice for capturing human behaviors.

Bayesian models of human cognition have met with a lot of success in recent years. But the use of rational models in cognitive science has been limited to specific cognitive processes in perception, memory and language processing [13]. Most of these rational models have been derived using Baye's rule, a particular theorem of probability that gives the relation between one conditional probability and its inverse. In other words, it relates the probability of a hypothesis given observed evidence and the probability of that evidence given the hypothesis. For example, if for a specific hypothesis H, $P(H)$ is the prior probability of H that was inferred before new evidence, E became available, we can calculate the posterior probability of H given E as

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \tag{1}$$

where $P(E)$ is the priori probability of witnessing the new evidence E under all possible hypotheses and $P(E|H)$ is the conditional probability of seeing the evidence E if the hypothesis H happens to be true (likelihood).

In the next section, we will propose the use of Baye's rule for our learning model and how it estimates the assumptions of the learner in the form of priori, likelihood and hypothesis space.

3 Multimodal Memory Game (MMG)

3.1 MMG Architecture

The multimodal memory game (MMG) provides a learning platform for the human players as well as the machine learner. The game architecture (Fig. 1) incorporates a user interface (game manager) that enables the human players to interact with the game and helps the machine learner to learn from the player's behaviour. The game involves two modalities of vision and language. There are two human players, i.e. the text to text generator T2T and the text-image to text generator TI2T.

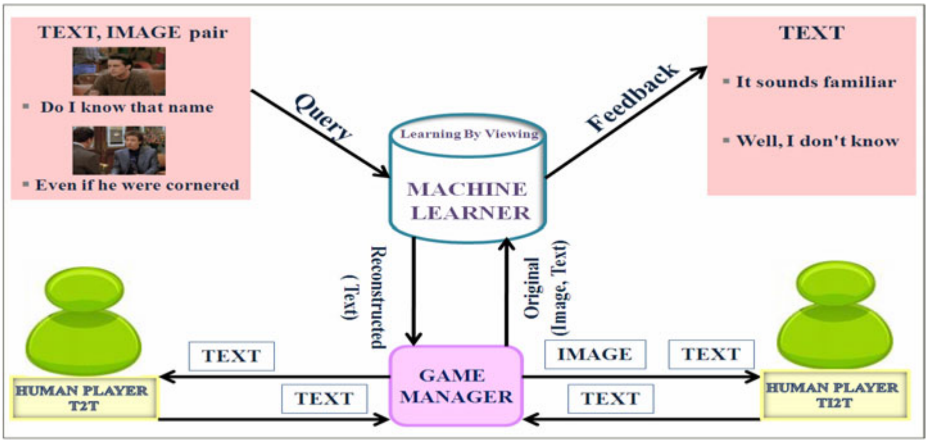


Fig. 1. Multimodal Memory Game Architecture

The game uses the multimodal information (image and text) from a video clip of a TV drama series, in this case, *Friends*. It proceeds in a quiz-like format. The game provide the human player with an interface for questions and answers, and the machine learner learns from the human input (and/or the correct answers which the game manager provide) while they play the game. We show the human players an episode of the selected TV drama after which they answer questions based either on a given text or text-image query taken from the video watched. We require the human player T2T to input the succeeding text for the queried text and similarly, the human player

TI2T provide the following text for the queried text-image pair. The game session continues in this manner until the user exits the game or answers all the questions required to attain a score that is recorded by the game manager.

3.2 MMG as a Cognitive Model of Language Learning

Recent research has provided ample evidence for use of Bayesian models of cognition to explain why humans behave as they do when presented with a specific task and the data to work with [14, 15, 16]. However, these models do not address human cognitive limitations in identifying what should be the optimal solution for the given data. They may use computational procedures that human learners cannot use.

In our proposed setup, the Bayesian learner seeks to learn from the language (incorporating the use of lexicon of words) input by the human player. Here the language acquisition is relatively easy for the Bayesian learner due to the fact that the human player has sound knowledge about the internalized structure of the observed data (English drama subtitles). We present the human player with some question $\boldsymbol{q}$, which is a segmented corpus of words and the player gives answer $\boldsymbol{a}$. This $(\boldsymbol{q}, \boldsymbol{a})$ builds the data $\boldsymbol{d}$ for machine learner. If we assume the human player to be the teacher and the machine learner to be the Bayesian learner in our framework, then the learner updates its hypothesis by the Baye's rule

$$P(h \mid d) \propto P(d \mid h)P(h) \tag{2}$$

where $P(h)$ is the learner's prior, $P(d|h)$ is the likelihood, and $P(h \mid d)$ is the posterior probability distribution of the hypothesis of the learner.

The Bayesian learner sees data produced by the teacher and forms a hypothesis about the approach used to produce that data. The learner then uses this hypothesis to produce data to make up its learning. For the Bayesian learner, the likelihood is 1 if the concatenated sequence of words in the hypothesis matches with those in the observed data. On the other hand, the likelihood is 0 if the sequence does not match the word sequence from the observed data.

In the above equation, we assume that the learner updates his belief with subsequent new examples and the teacher provides the learner with examples that tend to increase the learning. If the learner assumptions are in accordance with the observed data representation, the hypothesis is probable to have high prior probability. Varying these prior probabilities can result in characterization of new models which then can be used for the learners in a constrained learning environment.

4 The MMG Game Interface

We provide a user-friendly interface to the player that is easy to use and operate. The game offers the players to watch a video from the TV drama, for example a 20 minute-long episode of *Friends*. After watching the video clip, the player proceeds to choose one of the learning scenarios to play the game. For T2T learning, given text question, the player inputs the desired text. In TI2T learning, a text along with its visual context appears as a question and the player replies with the following text. Fig. 2 shows procedure for playing MMG games that gives an idea of how the game manager works.

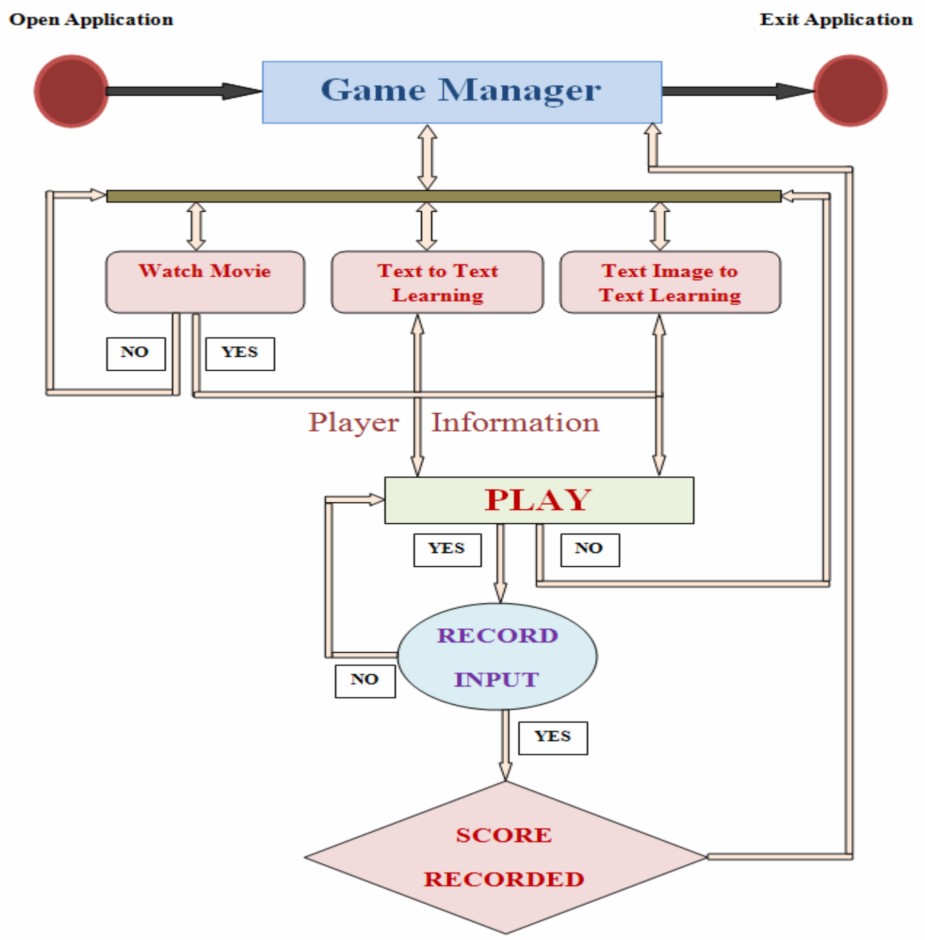
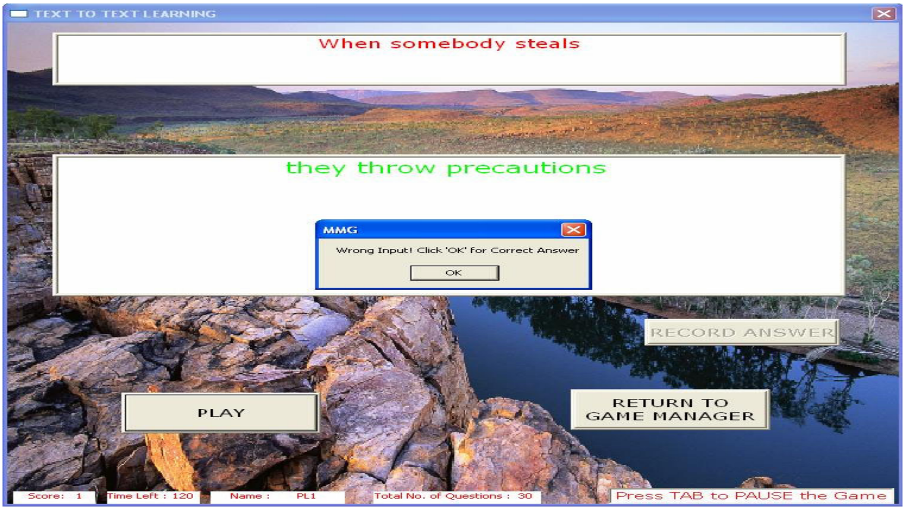


Fig. 2. Procedure for playing MMG games

When the human player clicks ‘PLAY’ to start the game, he provides personal information e.g. name, age and number of questions he would like to attempt during the game session. The players can choose the number of questions according to the need; they can be varied for multiple sessions or for longer sessions where the learning requires improvement. In case of text-to-text learning (T2T), a text from the TV drama episode appears on the screen and the player has to input the answer for queried text. The game interface for the text-to-text learning is shown in Fig. 3.

Similarly, in case of text-image-to-text (TI2T) query, an image along with text (subtitle) from the video clip appears on the screen. The player has to write the succeeding text to proceed to the next question as shown in Fig. 4. To make the learning process fast, the game time is managed and the players use a time quanta of 120 seconds for every question.



When somebody steals

they throw precautions

Wrong Input! Click 'OK' for Correct Answer

Score: 1

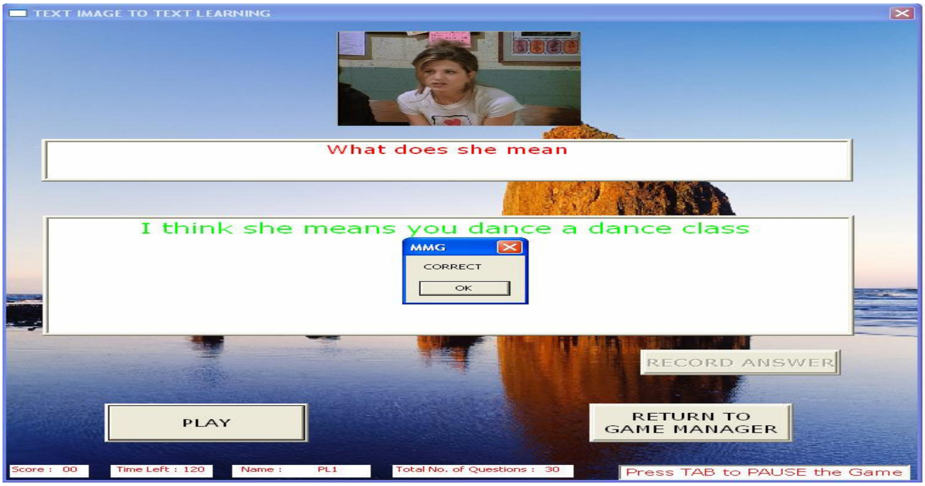
Time Left : 120

Name : PL1

Total No. of Questions : 30

Press TAB to PAUSE the Game

Fig. 3. User interface for text-to-text learning



What does she mean

I think she means you dance a dance class

CORRECT

Score : 00

Time Left : 120

Name : PL1

Total No. of Questions : 30

Press TAB to PAUSE the Game

Fig. 4. User interface for text-image-to-text learning

During this time duration, if the player does not input the answer, the game manager proceeds to the next question. The game manager provides the user with the facility to pause the game by keyboard stroke in case he is interrupted by some external means. It updates the player status as the game proceeds along with the time left for the question on the status bar at the bottom of the interface. When the session ends, the game manager stores the player information with the number of correct answers in a text file for further use and reference.

5 Experimental Setup and Results

The data set for the experiments consists of the image-text pairs taken from an episode of the *Friends* series. We performed experiments for observing text-to-text learning and text-image-to-text learning obtained by the human players during the game play. The five volunteers were asked to watch the 23 minute video clip and play 10 sessions for each version of the game. The game manager scores sessions individually for all the players. We use the parameter settings for our experiments as given in Table 1.

Table 1. Parameters for the MMG experiments

Total no. of texts/images	No. of query texts/ images per session	No. of sessions	Time for each question (sec)
294	30	10	120

For scoring the experiments, we set a 60% threshold for the player input to be mark as correct based on the comparison of count of words that match correctly with the correct sequence of words.

Table 2. Comparison of player input to correct sentence

Player input	Correct sentence	Threshold (%)
I would say	I would have to say	60

Threshold in Table 2 is calculated using the below relation

$$\text{Threshold (\%)} = \frac{\text{Number of Words from Player Input}}{\text{Number of Words in Correct Sentence}} \tag{3}$$

The player input will be marked as correct if the ratio of the input words sequence to the correct word sequence matches the desired threshold. The result of each game session is used to calculate the accuracy of the human recall. We define recall accuracy as the ratio of the number of questions attempted correctly to the total number of questions in a given session as shown in below equation;

$$\text{Recall Accuracy} = \frac{\text{Number of Correctly Answered Questions}}{\text{Total Number of Questions}} \tag{4}$$

We use the average of the calculated accuracy for the five volunteers and plot the accuracy values against the number of sessions attempted. In the first experiment, we provide textual context as the cue and ask the players to respond with the succeeding text. In the second case of text-image-to-text learning, we add image context along with the textual context of the video and asked the players to reply with the succeeding text for the queried image-text pair to observe the impact of image whether it aids in guessing the succeeding text. Similar parameter values are set for the text-image-to-text learning.

The volunteers play the game for 10 sessions and their scores are plotted against the number of sessions. The average accuracy curves for text-to-text learning and text-image-to-text learning in Fig. 5 shows a significant improvement in the human learning process from 5% to almost 60% during the 10 game sessions. The average accuracy curve for text-image-to-text learning produced a better response as compared to the text to text only case, however, the overall learning response for both the cases seem to have improved with the increasing number of game sessions.

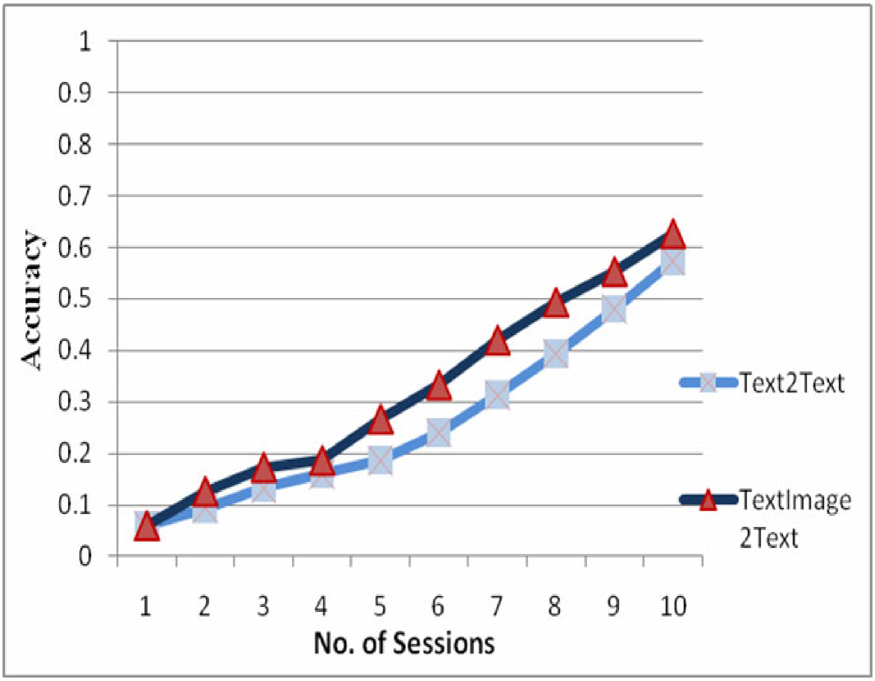


Fig. 5. Average of recall accuracy for text-to-text and text-image-to-text learning

Improvement in human learning can generally be achieved more rapidly by varying the game parameter settings, i.e. by exposing the human player to a larger portion (50 examples) of the training set (294 examples) in a single session rather than 30 questions may probably result in faster improvement (in term of the number of sessions) in learning response for the humans.

6 Concluding Remarks

We have presented an implementation of the multimodal memory game platform that can be used for studying the learning behaviors of humans and machines. Here we proposed a simple cognitive model of learning that can be used for characterizing human learning. We anticipate the proposed model approach will be helpful in answering a few deep questions about human cognition; e.g. how human mind makes predictions from a limited observed data and what specific forms of knowledge do they acquire that support human inference in many different tasks.

From our experimental findings, we show that the human learning response improves gradually with the number of game sessions for both scenarios of text-to-text learning and text-image-to-text learning. We observed an apparent improvement in the learning curve for text-image-to-text learning which establishes our belief that addition of image is helpful in improving the learning. With these results, we believe it's too early to conclude about the general learning tendency of the humans, especially in our experimental setting where the player exposure is only to a small portion of the entire training set in each session.

We believe that this learning setup could form an inductive learning model of cognition and can be utilized to achieve human-like machine learning. Adoption of this procedure for machines may lead to life-long learning without complaints and it will be interesting to see how the human and machine learning performances compare for a larger scale experiments.

By setting the experimental parameters in a cognitively more plausible way, the model can be used for different learning platforms. These may include e-learning or robot learning in which the interaction is based on several different modalities of information in a changing environment.

Acknowledgment

This work was supported in part by IT R&D Program of MKE/KEIT (KI002138, MARS), NRF Grant of MEST (314-2008-1-D00377, Xtran), the BK21-IT program of MEST and the Higher Education Commission of Pakistan.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
2. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach, 2nd edn. Prentice Hall, Englewood Cliffs (2002)
3. Griffiths, T.L.: Connecting Human and Machine Learning via Probabilistic Models of Cognition. In: Technical Program. 10th Annual Conference of the International Speech Communication Association (2009)
4. Zhang, B.-T.: Cognitive learning and the multimodal memory game: Toward human-level machine learning. In: IEEE International Joint Conference on Neural Networks, pp. 3261–3267 (2008)
5. Jensen, F.V., Nielsen, T.: Bayesian Networks and Decision Graphs. Springer, New York (2007)
6. Vlontzos, J.A., Kung, S.Y.: Hidden Markov Models for Character Recognition. IEEE Trans. Image Processing 1(4), 539–543 (1992)
7. Richardson, M., Domingos, P.: Markov Logic Networks. Machine Learning 62, 107–136 (2006)

8. Sutskever, I., Hinton, G.E.: Deep Narrow Sigmoid Belief Networks are Universal Approximators. Neural Computation 20, 2629–2636 (2008)
9. Zhang, B.-T.: Hypernetworks: A Molecular Evolutionary Architecture for Cognitive Learning and Memory. IEEE Computational Intelligence Magazine 3(3), 49–63 (2008)
10. Maragos, P., Potamianous, A.: Multimodal Processing and Interaction Audio, Video, Text. Springer Science Media, Heidelberg (2008)
11. Benczúr, A., Bíró, I., Brendel, M.: Cross-Modal Retrieval by Text and Image Feature Biclustering. In: CLEF 2007: Proceedings of Cross Language Evaluation Forum (2007)
12. Zhang, R., Zhang, Z., Li, M., Ma, W.-Y., Zhang, H.-J.: A Probabilistic Semantic Model for Image Annotation and Multi-Modal Image Retrieval. Multimedia Systems 1, 27–33 (2006)
13. Oaksford, M., Chater, N.: Ten Years of the Rational Analysis of Cognition. Trends in Cognitive Science 3, 57–65 (1999)
14. Griffiths, T.L., Tenenbaum, J.B.: Structure and Strength in Causal Induction. Cognitive Psychology 51, 354–384 (2005)
15. Pearl, L., Goldwater, S., Steyvers, M.: How ideal are we? Incorporating human limitations into Bayesian models of word segmentation. In: Proceedings of the 34th Annual Boston University Conference on Child Language Development. Cascadilla Press, Somerville
16. Goldwater, S., Griffiths, T., Johnson, M.: Distributional Cues to Word Boundaries: Context is Important. In: BUCLD 31: Proceedings of the 31st Annual Boston University Conference on Language Development, pp. 239–250. Cascadilla Press, Somerville (2007)