Technical report “Common sense reasoning through imagination”

S. Voronov
rdkl.hrd@gmail.com

To build algorithms for automatic problems solving, one often models human notions about problem solving procedure.

In this paper we construct an algorithm that answers questions about the spatial structure of objects, actions of which are described in the input text. The novelty of the algorithm lies in generation of additional visual information, just as a person responds to this type of questions, imagining spatial scenery inside brain. A sequence of frames is generated according to the similarity to the sentences of the source text. We use the three step answer generation procedure. The first step is the generation of all possible answers. The second step is the matching them with the obtained visual information. The third step is to select the answer, which has the maximum answer-video likelihood.

To generate visual information, that would be the most similar to the obtained set of models, the random generator of objects trajectories, based on Metropolis-Hastings algorithm, is used. The product of likelihoods of the most probable path through word model states is used as quality function of obtained video.

The constructed system can answer a set of questions about the time-varying spatial structure of objects, which are described in text form.

Keywords: *question answering; video generation*

Технический отчет “Вопросно-ответная система с автогенерацией пространственных структур”

С. О. Воронов
rdkl.hrd@gmail.com
Московский физико-технический институт, ФУПМ, Кафедра интеллектуальных систем

Моделирование человеческого подхода к решению той или иной задачи часто используется для построения автоматических алгоритмов. В работе предлагается алгоритм, способный отвечать на вопросы, связанные с пространственным расположением объектов, действия которых описаны в исходном тексте. Новизна алгоритма заключается в том, что для определения ответа производится генерация последовательности кадров о расположении объектов в пространстве и времени, аналогично тому, как это представляет себе человек.

Для генерации последовательности кадров, наиболее точно описывающих полученный набор моделей слов, используется генератор случайных траекторий объектов, основанный на алгоритме Метрополиса-Гастингса. Функцией качества полученной последовательности кадров является произведение правдоподобий моделей слов на наиболее вероятном распределении их состояний по кадрам.

Построенная система в состоянии отвечать на набор вопросов о изменяющихся во времени пространственной структуре объектов, изначально описанных в текстовой форме.

Ключевые слова: *вопросно-ответные системы; генерация видео*

1 Введение

How typical Question Answering system works? Firstly, it does question analysis. Secondly, system makes query for its database (it could be stored outside the whole system, like Internet). And the last step is to rank answers (to return the best one). Today, many people use AI assistants (like Siri or Google Now) in their daily routine. But these systems can only extract information from world, but they can not imagine your movements from natural speech. You might need it for help with some road problems in case you’re in trouble and want to find arguments for dispute. This paper provides approach that helps machine to understand objects movements from text (that movements would be generated) and to answer some questions about that.

According to [3], typical QA systems could be divided into three main categories:

1.1 Linguistic approach

The first QA systems (1960s) were NL query front-ends for knowledge database (like BASEBALL [5]). Database size made a limitation: these systems were able to answer questions inside restricted area. Next step of evolution was to acquiring Internet as knowledge database (examples: START [8, 9], [2] and [12]).

1.2 Statistical approach

Size of data, created by mankind, has increased importance of statistical techniques. They have been applied to the different stages of a QA system (analysis questions type, predictions about expected answers, etc). Famous systems are IBM’s statistical QA [7], [13]. According to [3], [1] has investigated the prospects of applying statistical methods to answer finding task in QA and discovered that these techniques performed quite well depending on the underlying data set characteristics – vocabulary size, the overlap between question and answers, between multiple answers, etc.

1.3 Pattern matching approach

Here one could use text patterns to answer some types of questions. For example, the question «Where is Disney located?» relies on pattern «Where is <object> located?» and produce answer «<Object> located at <Location> ». Many of the QA systems automatically recognize such text patterns from text passages rather than employing complicated linguistic tools to text for retrieving answers.

Idea of extending text QA to work with some types of media is pretty natural: there is a lot of information in the Internet in media form. In [14] authors try to enrich text answer with media (image or video) information from internet. The paper [6] tries to work only with key shots from video. The study [4] used NLP to create similar questions and rerank obtained videos queues with video analysis.

Some papers provided idea of answering questions about images or videos. For example, in [10] authors built Spot – system, that is able to answer specific questions about surveliance videos. START translated English question into inner representation (video filter) that used to transform (dynamically) raw video [8]. Also, neural architectures were used for QA about images [11].

This paper solves the following problem: to answer questions about objects movements by generation video-like info, describing these movements. Thus, we provide an approach for QA that is similar with human thinking: generation of video-like scene while answering question. We use idea proposed in [16] about words representation as multistate FSM (hidden Markov model) over features extracted from video. As [17] we represent all words as HMM, not only verbs. Using START [8] we construct sentence-specific structure from HMMs, representing separate words, considering the relations between the objects. Following [17], next we create a function $S : (B, s, \lambda) \to (\tau, J)$, where B represents the information extracted from a video clip, s represents the sentence, λ represents word meanings, τ is the video-sentence score, and J is a collection of tracks, one for each participant in the event described by the sentence, corresponding to the best video-sentence score. After we describe track generation procedure for given sentence. In the Section we use this generation procedure to answer hypothetical spatial questions.

2 Main section

2.1 Scoring

Every word from the original text is represented as a linear hidden Markov model. For each state of the each word model there is a prior distribution of features values (features depends on objects in a frame), which is responsible for the likelihood for (video frame, state of word model) pair.

Assume we have translation the function

$$tr : \{1, ..., n\} \to \{1, ..., s_m\}, \text{ such that } x > y \Rightarrow tr(x) \geqslant tr(y).$$

The function tr maps number of frame into assigned model state. Then one can compute likelihood:

$$\log(L) = \sum_{t=1}^{n} h(k^t, fr_t) + \sum_{t=1}^{n} a(k^{t-1}, k^t), \tag{1}$$

where k^t means state $tr(t)$, $h(k^t, fr_t)$ – how well state fits frame detections, and $a(k^{t-1}, k^t)$ – transition $k^{t-1} \to k^t$ cost.

Track detection part for video B uses the following equation:

$$\max_j \left(\sum_{t=1}^{T} f(b^t_{j^t}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}}, b^{t-1}_{j^t}) \right), \tag{2}$$

where $b^t_{j^t_l}$ – detection chosen for track (trajectory) l on frame t.

This equation could easily be solved by Viterbi [15] algorithm (in fact, dynamic programming). That will produce a lattice size $J \times T$, where J is maximum between number detections on each frame.

In this paper

$$f = \min(1, \max(-1, detection\_score)), \text{ and}$$
$$g = -distance(b, b') + optical\_flow(b, b').$$

Then we replace one model by many models, that are united into on big model with states equal to vectors of previous models states and solve these equation together (to find tracks fits this hyper model well):

$$\max_{\mathbf{J},\,\mathbf{K}} \left[\sum_{l=1}^{L} \left(\sum_{t=1}^{T} f(b^t_{j^l}) + \sum_{t=2}^{T} g(b^{t-1}_{j_l^{t-1}}, b^{t-1}_{j_l^t}) \right) + \sum_{m=1}^{M} \left(\sum_{t=1}^{n} h(k^t_m, fr_t) + \sum_{t=1}^{n} a(k^{t-1}_m, k^t_m) \right) \right] \tag{3}$$

2.2 Generation

We want to generate visual information that would be similar to the simplest videos. That means here will be objects and their movements. But in real videos (with small amount of exceptions) object will move slowly through frames (one might think about video as sequence of frames). Therefore, if we interpolate trajectory of any object through space and time it will be mostly smooth. Lets try to generate something, that would be similar to real trajectories. We will use process of random search of available trajectories space. Due to huge dimensionality of this space there is a necessity of limitation. It produces the following requirement: it would be good if whole trajectory could be controlled by relatively small amount of parameters. That is why B-splines as trajectories were used.

```
Generate initial track $t$;
while iteration number is not reached do
    Generate track candidate $t'$;
    $\alpha \leftarrow \frac{s(t')}{s(t)}$, where $s$ is track score (in the common case from Equation ??);
    if $\alpha > 1$ then
        replace $t$ with $t'$
    else
        accept the candidate with probability $\alpha$
    end
end
```

Algorithm 1: Metropolis-Hastings track(s) generation

2.3 Question answering

Due to relatively small types of questions could be asked about object positions, we use pattern matching approach (from Section 2).

We asked the following questions:

– Who is near *obj* in time t?
– How far is obj_1 from obj_2 in time t?
– Where was *obj* in time t?
– Is obj_1 left of the obj_2 in time t?
– Is obj_1 right of the obj_2 in time t?

To improve this algorithm we generate 3-4 different scenes of track and use them all. In this case we could use Cosine similarity or correlation between answers scores vector and initial text scores vector for ranking. But this approach does not achieve appropriate quality with complex sentences. Also, this could not work with questions about exact time. So we used the following scheme.

1. Generate 3-4 video-like media
2. Parse question using pattern approach and produce answers
3. Use absolute positions of objects in video to find the best answer

3 Conclusion

This paper solves the following problem: to answer questions about objects movements by generation video-like info, describing these movements. We provided an approach for answering about spatial questions, that uses video-like information generation similar with human thinking. We proposed two algorithms for objects tracks generation (where objects are taken from input sentence). Also we proposed two algorithms for answers ranking (comparing sentences and direct video analysis). We achieved next results: 74% questions about two input objects have been answered correctly. Also, 32% of questions about sentences with three objects have been answered well. In the future this system could be a part of QA system (and produce additional information for some spatial questions).

Future work: try other types of generation algorithms, other types of tracks (instead of B-splines) and speed up algorithm.

Литература

[1] Berger A, Caruana R, Cohn D, Freitag D, and Mittal V. Bridging the lexical chasm: statistical approaches to answer-finding. In *23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192– 199, 2000.

[2] H. Chung, Y.I. Song, K.S. Han, Yoon D.S., J.Y. Lee, and H.C. Rim. A practical qa system in restricted domains. In *42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004.

[3] Sanjay K Dwivedi and Vaishali Singh. Research and reviews in question answering system. In *International Conference on Computational Intelligence: Modeling Techniques and Applications*, 2013.

[4] Lei Gao, Guangda, Yan-Tao Zheng, Richang Hong, and Tat-Seng Chua. Video reference: A video question answering engine. In *17th International ACM Conference in Multimedia*, 2009.

[5] B.F. Green, Wolf A.K., Chomsky C, and Laughery K. Baseball: An automatic question answerer. In *Western Computing Conference*, volume 19, page 219–224, 1961.

[6] Richang Hong, Jinhui Tang, Hung-Khoon Tan, Chong-Wah Ngo, Shuicheng Yan, and Tat-Seng Chua. Beyond search: Event-driven summarization for web videos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM*, 7, 2011.

[7] A. Ittycheriah, M. Franz W.J. Zhu, A. Ratnaparkhi, and R.J. Mammone. Ibm's statistical question answering system. In *Text Retrieval Conference TREC-9*, 2000.

[8] Boris Katz. Annotating the world wide web using natural language. In *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet*, 1997.

[9] Boris Katz, Gary Borchardt, and Sue Felshin. Natural language annotations for question answering. In *Proceedings of the 19th International FLAIRS Conference*, 2006.

[10] Boris Katz, Jimmy Lin, Chris Stauffer, and Eric Grimson. Answering questions about moving objects in surveillance videos. In *AAAI Technical Report SS-03-07*, 2003.

[11] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. arXiv, Oct 2015.

[12] A. Mishra, N. Mishra, and A. Agrawal. Context-aware restricted geographical domain question answering system. In *IEEE International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 548–553, 2010.

[13] A. Moschitti. Answer filtering via text categorization in question answering systems. In *15 th IEEE International Conference on Tools with Artificial Intelligence*, pages 241–248, 2003.

[14] Liqiang Nie, Meng Wang, Zhengjun Zha, and Guangda Liand Tat-Seng Chua. Multimedia answering: enriching text qa with media information. In *34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011.

[15] A. J. Viterbi. Error bounds for convolutional codes and an asymtotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 1967.

[16] J. Yamoto, J. Ohya, and K. Ishii. Recognizing human action in time-sequential im- ages using hidden markov model. In *IEEE Conference on Computer Vision and Pattern Recognition*, page 379–385, 1992.

[17] H. Yu, N. Siddharth, A. Barbu, and J. M. Siskind. A compositional framework for grounding language inference, generation, and acquisition in video. *Journal of Artificial Intelligence Research*, 52:601–713, 04 2015.