Semantic Parsing of Action Text for Text-to-Scene Conversion

Fuping Yang, Linna Wang*
College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065, P. R. China
wanglinna1027@163.com

Abstract. Understanding the action text is the main problem of dynamic scene generation in text-to-scene conversion. However, the existing basic research about it is still lacking, and its semantic parsing method is not perfect. Therefore, based on the idea of action chain structure, the formal expression of it is established on the basis of defining action text, classifying action text and analysing the features of action text. The method of action text semantic parsing is proposed combing with the action text semantic dictionary. The experimental results show that the method completes the semantic parsing for the action text to a certain extent and provides a reference for computer to understand the semantic of action text.

1. Introduction
Nowadays, human want to express their thoughts through visual images or 3D scenes, instead of using words. The text-to-scene conversion is generated under such a demand, and it can convert the visible natural language description into its corresponding 3D static scene or animation through computer. The application areas of text-to-scene conversion are very wide, including education, housing design, traffic accident scene recovery, etc. Text-to-scene conversion is an emerging discipline and it was proposed by B.Coyne and R.Sproat of AT&T [1]. There is no uniform definition of basic concepts and corresponding international consensus about it[2]. But there are still many experts and scholars from all over the world exploring it. The WordsEye[3-4] developed by AT&T in the United States is based on a vocabulary knowledge base and 3D model library used to recognize and process text scenes, but only for static scenes. Stanford University researched the Stanford CoreNLP[5] to processing the language. Chang A[6-7] and others developed a scene description platform and invited a large number of people to describe the given scene in text, then established a description corpus corresponding to the scene. On this basis, the SceneSeer[8] was developed, which uses verb as the core and uses simple text to interact to generate scenes. Chinese Academy of Sciences has developed the swan system[9] which can automatically generate animation, but requires a huge grammar rule base and knowledge base. Harbin Institute of Technology[10] and Chongqing University of Posts and Telecommunications[11] also conducted in-depth theoretical research on spatial layout, entity attribute, et al, then proposed effective technical solutions.

2. Action text semantic parsing
2.1. The definition and classification of action text
In text-to-scene conversion, action text is generally a visual text that reflects the semantics of entity’s action. Due to the ambiguity and flexibility of natural language, action text refers not only to verbs, but also to nouns or other words. Therefore, the action text is defined from two perspectives: first, it
appears as verb in the text and describes the specific behaviour of the entity. Second, it appears as non-verb in the text, but it contains action semantics. For example, "games" is a noun in "The Games Are Exciting", but its semantic contain the specific actions: running, swimming, long jump, and so on. Therefore, the “game” is the action text.

Verbs need to be classified before parsing it. Different from the existing verb classification methods, the verbs are classified from the perspective of visibility in order to adapt the needs of text-to-scene conversion. First, according to the semantic characteristics and different visualization methods of verbs, they are divided into two categories: direct visualization verbs and indirect visualization verbs. Secondly, direct visualization verbs are divided into action text and non-action text. The following sections focus on the action text.

The direct visualization verbs are divided into meta action, compound action and complex action according to the complexity of the action text. Meta action does not need to be decomposed again. From the perspective of semantics, it has independent semantics; from the perspective of technical implementation, the existing technology has encapsulated some simple meta actions into packages that can be used directly when generating scenes. Such as run, jump, etc. Compound action is an ordered combination of meta action. For example, "high jump" can be broken down into “run-up-take-back-lading”. The semantic of complex action is more complex than compound action. It contains six different combinations: "meta action + compound action", "complex action + compound action", "meta action + complex action ", " complex action + complex action ", " complex action + complex action ", " meta action + compound action + complex action ". The following sections will focus on the meta action and compound action. The above verb classification is shown in Figure 1.

2.2. Features and structure of action text

The features of action text can be summarized from two perspectives: semantic features and performance features.

The semantic features of meta-actions are reflected in two aspects: irreducibility and smallest action unit. Irreducibility means that the meta actions can be directly understood and processed by computers. Because of this feature, the semantic parsing of compound actions can be based on meta actions, that is, meta actions are the smallest unit of action. The performance feature refers to that the meta actions mostly appear as monosyllabic verbs or "monosyllabic verbs + entity". The semantic of the same compound action is different with different objects. The performance feature is that they usually appear in the form of "verb + noun" and "verb + verb".

The "ordered combination" of the action text mentioned is the action chain structure. The action chain of a meta action is itself, and the action chain of a compound action is an ordered combination of several meta actions that can express its semantics. The action chain structure has three characteristics: diversity, granularity relativity and effectiveness. Diversity means that the same action text may have different action chains due to different objects. Granularity relativity means that different context and importance of action texts result in the same action text and objects need to be characterized with different granularities. Effectiveness means that as long as the action chain can show the semantics of the action, it is said to be the effective chain of the action text. Different people have different understandings for the same text. Therefore, this paper will focus on the effective action chains. In summary, the action chain can be expressed as follows:
$$T_a = <\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_n>$$  \hspace{1cm} (1)

where $T_a$ is action text, $<\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_n>$ is the action chain corresponding to $T_a$. When $T_a$ represents meta action:

$$T_a = <a>$$  \hspace{1cm} (2)

where $a$ is meta action, that is, $\bar{A}_i = a_i, \bar{A}_2 = a_2 = \cdots = a_n = 0$. When $T_a$ represents compound action, $\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_n$ is the set of corresponding action chains, $\bar{A}_i = (a_{i1}, a_{i2}, \ldots, a_{in})^T$ ($1 \leq i \leq n, m \in N^+, N^+$ is positive integer) represents one of the action chains. $a_{i1}, a_{i2}, \ldots, a_{in}$ is several meta actions that make up the action chain.

2.3. Construction of action text semantic dictionary

The four types of factors that affect action text semantic are action type, object categories, action methods and action effects, which are expressed as follows:

$$T_a = \ll C_a, C_e, M_a, F_a \gg$$  \hspace{1cm} (3)

where $T_a$ is action text, $C_a$ is action category, $C_e$ is object category, $M_a$ is type of action, $F_a$ is action effect.

The semantic dictionary of action text is built around $M_a$ and refers to Modern Chinese Classification Dictionary, it can be divided into four categories: head movement, upper limb movement, lower limb movement, and body movement. The dictionary consists of five layers. The first layer contains upper limb movements, lower limb movements, head movements and body movements. The second layer and the third layer are further subdivided into 29 sub-classes and 106 sub-classes. The fourth layer divides synonymous and homogeneous and the fifth layer is specific action text. The dictionary structure is shown in Figure 2.

![Semantic Dictionary of Action Text](image)

Figure 2. The structure of action text semantic dictionary.

The dictionary contains nearly 3000 meta actions and compound actions. The action chain sets are constructed for different types of compound actions. In order to ensure semantic integrity, the "longest action chain" is used as a standard. The action chain sets corresponding to any kind of action text are represented as follows:

$$S_i(h) = \{L_i \mid i \in N^+\}$$  \hspace{1cm} (4)

where $S_i$ is the set of action chain, $h$ is compound action, $L_i$ is one of the action chain of $h$, $N^+$ is positive integer.

The process of constructing the action chain set is as follows: first, crawling different types of 10,000 articles from the Internet, such as diaries, fairy tales, etc. Each contains around 1,000 words. Second, for each word in the dictionary, find it in the corresponding sentence in the corpus. Then, manually construct its corresponding action chain structure, extract related objects, and finally mark it...
as a meta action or a compound action. There are more than 1,300 longest action chains established through the above steps. Theoretically, the number of action chains can be infinite, but the actual requirement is making the computer understand the semantic of the compound action text. Therefore, the limited action chains will be focused.

2.4. Marking system and parsing process of action text
First, the text is pre-processed, then marking them according to Table 1.

| Type of word               | Symbol |
|----------------------------|--------|
| Direct visualization verb  | $V_d$  |
| Indirect visualization verb| $V_i$  |
| Action text                | $T_a$  |
| Non-action text            | $T_{na}$|
| Meta action text           | $T_{ma}$|
| Compound action text       | $T_{ca}$|
| Complex action text        | $T_{da}$|

Matching the action text with the dictionary. If match successful, $S_w = \{T_w \mid 1 \leq i \leq c\}$ is be formed. $T_w$ is the corresponding word in the dictionary, $c$ is the total number of words in dictionary. If multiple words match successfully, use the dictionary's classification structure to filter out the words which do not match the action semantics in the text. If match fails, use the trained Word2Vec to convert the word to vector for cosine similarity calculation. After that, the top m action texts which are most similar among them are selected to form a similar word set. Finally, the most similar words are selected as the corresponding word of the action text in the dictionary. The process is shown in Figure 3.

3. Experiment

3.1. The data of experiment and evaluation indicators
In the experiment, "the eighth radio gymnastics" was selected as the test corpus. The word vector in the similarity calculation is obtained by using the trained 64-dimensional Word2Vec model. The corpus used in the model total 32G and includes Wikipedia, news, diary, novel, etc. Finally, use the precision, recall, and F-measure to evaluate the experimental results.

3.2. Experimental results and improvements
Firstly, the test corpus is pre-processed. Then extracting verbs and automatically generating an "action file". Action texts that are not marked as verbs but have action semantics are manually added to the action file. There are 244 action texts in the test corpus. Two methods based on a dictionary, combining dictionary and similarity is used in experiment. The top 20 words with the highest
similarity are calculated, and the words which are most similar to the action semantic in the text are selected. The experimental results and effects are shown in Table 2.

| Method                           | Precision | Recall | F-measure |
|----------------------------------|-----------|--------|-----------|
| Based on the dictionary          | 99.06%    | 43.03% | 60.08%    |
| Combining dictionaries with similarity | 91.87%    | 54.92% | 68.72%    |

It can be seen in Table 2 that the combination of the dictionary and the similarity has a higher F-measure, but the recall of both methods is lower. After analysis, the reason is some action texts cannot completely correspond to words in the dictionary. In addition, and the semantic expression of each word in dictionary is more specific, which is different from the expression in the actual text. However, it is found that the semantic text of these unrecognized actions is similar to some words in the dictionary. By analysing the text features of the action text and the corpus, it can be known that the reason for this phenomenon is that in the actual text, more actions and entities appear in combination. Therefore, the second method is improved. When calculating the similarity, the action and the entity are combined, that is, ‘n+v’ is extracted as the action text, and the similarity is matched with the words in the dictionary. The improved experimental results and effects are shown in Table 3.

| Method                           | Precision | Recall | F-measure |
|----------------------------------|-----------|--------|-----------|
| Based on the dictionary          | 99.06%    | 43.03% | 60.08%    |
| Combining dictionaries with similarity | 92.06%    | 80.74% | 86.03%    |

It can be seen in the table 3 that after the improvement, the accuracy, recall, and F-measure of the second method have been improved to a certain extent compared with before. Therefore, this method is be selected. Through the improved method, there are more 17 compound action and 6 action chains obtained through the dictionary, 9 action chains obtained through the text structure. The effect comparison before and after improvement is shown in Figure 4.

![Figure 4. Comparison of experimental results before and after improvement.](image)

4. Conclusion

Aiming at the actual needs of action text understanding in context transformation, the method for semantic parsing of action text is proposed. The above work defines the related concepts about action texts and classify them, then analysing the semantic features of them. Constructing a formal expression of the action chain structure. In addition, combining the semantic dictionary of action text and similarity calculation to conduct semantic parsing experiments. Finally, improving the experiment after analysing the feature of action texts and corpus. The experiment proves that this method achieved the semantic parsing of action text to a certain extent. For the problem of polysemous semantic ambiguity, it can also be judged by the structure of the dictionary.

At present, the research on action texts in context transformation is still in the exploratory stage. The research in this paper also needs to be improved. Further research will focus on the following
aspects: first, automatically filtering the most similar action text by combining the goal word and the context from the set of similar word. Second, for action text with multiple action chains, they are screened out based on their characteristics such as agent, subject, and action category, the structure of the action chain in accordance with the current semantics.

References
[1] Coyne B, Sproat R. (2001) WordsEye: An automatic text-to-scene conversion system. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. United States: ACM. 487-496.
[2] Lee H, Chang A, Peirsman Y, et al. (2013) Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics, 39(4): 885-916.
[3] Rouhizadeh M, Bauer D, Coyne B, et al. (2011) Collecting spatial information for locations in a text-to-scene conversion system. In: Proceedings of the Workshop on Computational Models of Spatial Language Interpretation and Generation(COSLI 2010). Netherlands: CEUR-WS. 16-23
[4] Rouhizadeh M, Bowler M, Sproat R, et al. (2010) Data collection and normalization for building the scenario-based lexical knowledge resource of a Text-to-Scene conversion system. In: Proceedings of 2010 5th International Workshop on Semantic Media Adaptation and Personalization(SMAP). United States: IEEE Computer Society. 25-30
[5] Manning C D, Surdeanu M, Bauer J, et al. (2014) The Stanford CoreNLP natural language processing toolkit. In: Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore: ACL.
[6] Savva M, Chang A X, Hanrahan P, et al. (2016) PiGraphs: learning interaction snapshots from observations. ACM Transactions on Graphics (TOG). 35(4): 139.
[7] Savva M, Chang A X, Hanrahan P. (2015) Semantically-enriched 3D models for common-sense knowledge. In: Computer Vision and Pattern Recognition Workshops. IEEE. 24-31.
[8] Chang A X, Eric M, Savva M, et al. (2017) SceneSeer: 3D scene design with natural language. arXiv: 1703.00050[cs.GR].
[9] Ru-Qian L U. (2002) From story to animation-full life cycle computer aided animation generation. IEEE/CAA Journal of Automatica Sinica (JAS). 28(3): 321-348.
[10] Li H J, Zhao T J, et al. (2005) Graphic retrieval based on limited semantics. In: The IEEE International Conference on Natural Language Processing and Knowledge Engineering. Wuhan. 535-539.
[11] Yang F, Zhou Y, Luo X. (2015) Scene layout in text-to-scene conversion. In: International Conference on Systems and Informatics. IEEE. 891-895.