Visibility Tagging of nouns in Text-to-Scene Conversion

Fuping Yang, Xi Cheng, Qi Chen

1School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China
2Trip.com Group, Shanghai, China

Abstract. The recognition and extraction of visual text information is the basis of the realization of Text-to-Scene Conversion, while nouns are the main expression form of entity, event, orientation, action and other information in text, but not all nouns are suitable for visualization, so how to recognize the visibility of nouns is the key problem of Text-to-Scene Conversion. Based on the definition and classification of noun visualization, this paper establishes a keyword visibility dictionary. Furthermore, in view of the static limitations of the dictionary, a dynamic expansion method based on semantic similarity is studied, and an optimization method for the visibility annotation of nouns is proposed in combination with unsupervised clustering. The experimental results show that the proposed method is effective and feasible for nominal visibility annotation of common texts, and the accuracy of the optimized method is 16% higher than that of the simple semantic similarity method.

1. Introduction

Text-to-Scene Conversion is also called visual natural language description. Its purpose is to help people understand the meaning of natural language text more intuitively. Natural language [1] is a simple and effective medium to describe the visual concept and mental image. It is a difficult and complex task to visualize it. In the process of image generation from natural language description [2], we should first consider the real world, and find the key visual information from the sentences representing the most basic concepts in virtual and real environment.

Text-to-Scene Conversion [3] involves two steps, natural language processing and scene visualization. Natural language processing is used for text processing, text understanding, then the visual information is identified and extracted. Scene visualization is used to visualize information in text and render scene. At present, some researchers at home and abroad have carried out relevant research on text-to-scene conversion. Language based layout system [4] is a rule-based spatial manipulation system established by cognitive linguistics, which operates objects by searching rule base. WordsEye [5] system is the first system to generate 3D scene through natural language description. The system strictly requires the format of input text. Chang A [6] proposes a 3D indoor scene generation method based on a large number of vocabulary. Swan system [7] uses the story written in natural language and the animation system generated by the whole process of computer-aided animation generation technology. In the above research, in the text processing stage, language rules are often used to identify and extract visual information. We should recognize and extract the visual information from the perspective of understanding the text. As the key expression of the visual information of the text, it
is essential to realize the visualization of the text and solve the visual analysis of the noun. This paper analyzes the visual of nouns in Chinese text, and gives the definition of the research, the classification and discrimination rules of the visual of nouns, and adopts the method of automatically tagging the visual of nouns based on dictionaries.

2. Visualization of Nouns
The visualization of nouns is to analyze and reason the nouns, confirm the visual of the nouns, parameterize them, and then through computer graphics and image processing techniques, the nouns are converted into graphics or images and displayed on the screen. From the perspective of visualization, due to the diversity of nouns, nouns can be divided into: visualization nouns, non-visualization nouns. Visualization noun refers to the noun that can be parameterized and transformed into corresponding semantic graph or image to be displayed on the screen through computer graphics and image processing technology. Such nouns usually have spatial or procedural characteristics, mostly representing entities, activities, actions, etc. Non visualization nouns, on the contrary, have no clear features and fuzzy boundaries, so it is impossible to find the parameters clearly when parameterizing them. Such nouns generally represent the potential logic and laws in nature.

According to the complexity of feature items and parameter values obtained during the noun visualization process, it can be divided into direct visualization nouns and indirect visualization nouns. Direct visualization nouns can directly determine the characteristics of nouns and the acquisition of corresponding eigenvalues. Indirect visualization nouns refers to the nouns in the visualization process that need to be able to find the visual features and eigenvalues that should be visualized by one or more inferences.

In summary, the following text will be used to indicate the following. $n$ represents all noun sets, $vn$ represents a collection of visual nouns, $nv$ represents a collection of non-visual nouns, $dvn$ represents a direct visualization of a noun set, $ivn$ represents a collection of indirect visual nouns.

3. Construction of Dictionaries
In this paper, we adopts the method of constructing visual noun dictionary to solve the problem of visibility tagging of nouns. First the keyword dictionary is constructed, and then the dictionary is expanded. Since there is currently no dictionary directly for nouns, the construction of the keyword dictionary is performed manually. The keyword dictionary construction step is shown in Fig 1.
3.1. Data Preprocessing
Taking the second edition of "Modern Chinese Dictionary" as the corpus, this article uses the jieba word segmentation tool to classify the corpus, part-of-speech tagging, denoising, and extracting nouns. The corpus has a total of 44,893 words. After the processing is completed, a noun set is obtained. The noun set has 25018 words. Then we divide the noun set visually according to the following rules.

3.2. Rules For The Visual Of Nouns
No matter direct or indirect visual terms, the visualization of noun will be carried out by the direct visualization noun. The visualization of nouns consists of two steps, one is the mapping of nouns to entities, and the other is the visualization of entities. From these two steps, the relevant division rules can be summarized.

The mapping of nouns to entities can be solved by the basic classification of nouns. From the correspondence between nouns and entities, nouns can be divided into the following two categories, entity nouns and non-entity nouns. The entity noun is the title of the entity existing in the world. According to the entity described by the entity class noun, dividing the entity. In addition to the category, the visualization of the entity is also considered. Therefore, the growth law and the shape structure of these entities are also needed. Distinguish from appearance texture [8]. The specific division is shown in Fig 2. In this paper, $EN$ is used to represent the set of entity nouns and $NEN$ is used to represent the set of non entity nouns.
The visualization of visual nouns is essentially the visualization of entities, so visual nouns can be further analyzed by analyzing the complexity of entity visualization. In the visualization of the entity, it is necessary to explore from the visible range and the physical structure of the human eye. The research shows that the smallest particle of the human eye is 0.1mm-0.2mm, and the wavelength of the visible wave of the human eye is between 400nm-760nm. And the human visible entity must have a certain color, shape or outline.

Entity and noun have the following mapping relationship:

\[ f : N \rightarrow E \]  \hspace{1cm} (1)

\( N \) : Represents a collection of nouns \( n \).
\( E \) : Represents a collection of entities \( e \).
\( e(c, a, v) \) represents the entity's triple structure, \( c \) represents the conceptual domain of an entity, \( a \) represents the attribute set of the entity, \( v \) represents a collection of attribute values.

\( a(str, tet, col, si) \) represents the attribute structure of the entity, \( str \) represents the structural properties of an entity, \( tet \) represents the texture property of the entity, \( col \) represents the color attribute of the entity, \( si \) indicates the size of the entity.

Each attribute has its corresponding attribute value, \( strv \) represents the structural property value of the entity, \( tetv \) represents the texture property value of the entity, \( colv \) represents the color attribute value of the entity, \( siv \) represents the size attribute value of an entity.

In summary, the rules for judging the visualization noun are summarized as follows:

- If \( strv > 0, tetv > 0, 0.400nm < colv < 770nm, siv > 0.02mm \), \( f \) is simple and \( n \in EN \), then \( n \in dvn \);
- If \( n \in EN \), \( str, tet, col, si \) have any value is unknown or \( si < 0.02mm, colv < 400nm, colv > 760nm \) and
\( f \) is simple, then \( n \in \text{ivn} \);
- If \( n \in \text{NEN} \) and \( f \) is complex, then \( n \in \text{ivn} \);
- If \( n \in \text{NEN} \) and \( f \) is non existent, then \( n \notin \text{ivn} \).

3.3. Dictionary Building Results

According to the above-mentioned noun visual division rules, the noun sets are manually divided into three categories: direct visual nouns, indirect visual nouns, and non-visual nouns. After classification, direct visualization nouns have 13346 words, indirect visualization have 10,800 nouns, and 872 non-visualized nouns.

4. Dynamic Expansion of Dictionaries

After the construction of the keyword dictionary, because the word set is static, there is a problem that new words cannot be marked when directly using the dictionary to mark the visibility of nouns, so this paper uses the way of recognizing the nouns in the text and expanding it to the dictionary to further improve the dictionary.

In this paper, 306129 articles are obtained from Wikipedia as the original corpus. Firstly, the corpus is simplified to remove other words except Chinese, and then we use the Jieba word segmentation tool to segment the corpus, mark the part of speech, and deal with the stop words to get a word set of 208539295 words. Next, we use word2vector to process the word set and get a word vector set. Because of the large amount of data, we delete the word vector outside the noun to get a noun vector set of 610744 word vectors, and randomly sample 1000 noun vectors as test samples to prepare for later experiments.

In this paper, we take two ways to expand the dictionary dynamically, one is based on semantic similarity, the other is based on the combination of unsupervised clustering and semantic similarity.

4.1. Extension Based on Semantic Similarity

Semantic similarity is the basic way to judge the relationship between word vectors, so this paper first uses semantic similarity calculation to judge the visual classification of unknown nouns, and the experimental steps are shown in Fig 3.

Cosine similarity is used to calculate the similarity. Given two vectors \( A \) and \( B \), cosine similarity \( \theta \) is given by point product and vector length, as shown in formula (2):

\[
\theta = \frac{A \cdot B}{\|A\| \|B\|}
\]
Select the top 10, 20, 30 and 40 words that are closest to the cosine similarity of the target words, calculate the frequency of the visual category of nouns in these phrases, select the category with the most frequency, judge it as the visual category of the target words, and then tagging the target words visually.

Fig 4 shows the statistics of accuracy, accuracy, recall rate and F1 value from the experiment of selecting similarity threshold. From the experimental results, when the number of words is 10, the accuracy is the best.

\[ \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \] (2)

4.2. Extension Based on Unsupervised Classification and Semantic Similarity

In theory, classification can be realized by supervised learning and unsupervised learning, but supervised learning requires a high coverage of classification features of self built dictionaries, while the data volume of self built dictionaries in this paper is limited, so it is of little significance to use supervised learning for classification. According to the above characteristics, unsupervised learning is used to train the whole word vector, and the optimal cluster set is combined with an experiment to form an optimization algorithm. The experimental steps are shown in Fig 5.

**Figure 4.** Threshold selection of noun similarity

**Figure 5.** Extension based on unsupervised classification and semantic similarity
In this paper, K-means clustering algorithm [9] is used for training. The training vector set is divided into m disjoint clusters, each of which is described by the mean value of the samples in the cluster. This mean is often called the centroid of a cluster. The k-means algorithm is designed to select a standard centroid that minimizes the sum of squares of inertia or sum within a cluster:

$$\sum_{t=0}^{m} \min_{\mu_i \in \mathcal{C}} \left( \| x_j - \mu_i \|^2 \right)$$

Before clustering the data, it is necessary to analyze the data. Because in very high dimensional space, Euclidean distance tends to expand. Therefore, the nonlinear t-sne dimensionality reduction algorithm is used to reduce the curse of dimensionality and speed up the calculation.

Use t-sne to reduce the set of self-built dictionary vectors to 2D and 3D and render as shown in Fig. 6 and Fig. 7.

Figure 6. 2D

Through observation, it is found that the feature clusters of different visual classifications are more obvious, among which visual nouns and non-visual nouns are the most obvious, and their features will be stripped and magnified in different dimensions.

In this paper, t-sne is used to reduce the dimension of the whole word vector, and the target dimension is set as 10. Then k-means algorithm is used to cluster the processed vector to get the training model.
In this paper, we propose an experiment of visual classification of subsequent nouns by combining cosine and clustering weight optimization formula. The formula is as follows:

$$ sim = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{1}{\cos} * KL \right)^{2}} $$

(4)

$sim$ is the similarity between the target word and the visual category, $N$ is the number of nouns in each visual category in the set with the best cosine similarity, $\cos$ is the cosine distance between the target word and the similar word, $KL$ is the proportion of the total clustering categories of the former similar word in the current set with the best cosine similarity, and the standard product of the two is the similarity of the visual category, the bigger, the better.

The cosine distance between the word vector result set trained by word2vec and all the words in the self built visual dictionary is calculated by using the 1000 noun test samples randomly sampled before. Select the top 10, 20, 30 and 40 words that are closest to the cosine similarity of the target words, use the cosine and clustering weight optimization formula to calculate the most similar visual category, judge it as the visual category of the target words, and mark the target words visually.

The comparison results of experiment A and experiment B are shown in Fig 8. The solid line is the experiment B based on the combination of unsupervised clustering and similarity, and the dotted line is the experiment A based on cosine similarity calculation.

![Figure 8. Comparison of experiment A and experiment B](image)

It can be seen from the experimental results that, compared with the direct use of cosine distance to compare vector similarity to obtain visual categories, the cosine and clustering weight optimization method improves the accuracy by 16%, the accuracy by 11%, and the recall rate and F1 value are significantly improved.

5. Experimental Effect
In order to verify the feasibility and effectiveness of the proposed method, this paper has carried out relevant experiments and achieved good experimental results. This paper uses argumentation, expository text, narrative text, prose, fairy tale five types of text, a total of 10, as experimental data. By using the Jieba word segmentation tool to segment the text, mark the part of speech, and then mark the nouns visually. The data set contains 1994 nouns in total. The accuracy of noun visibility annotation is 85% by experiment and artificial judgment.

6. Conclusion
Noun is the expression of key visual information in text, so it is necessary to analyze it visually. In this paper, the visual annotation method of nouns is studied, and the method studied in this paper is applied to the visual annotation of nouns in common texts. In the second part of this paper. In the second part of this paper, the visual definition and classification of nouns are proposed. Then, we propose a dictionary way to mark nouns visually. According to this method, the detailed design and
implementation of dictionary construction are described. Based on the construction of key words dictionary, we use the method of unsupervised clustering and semantic similarity calculation to expand the dictionary dynamically, and then realize the visual annotation of nouns. The experimental results show that this method is feasible and improves the accuracy of visual noun tagging. In the future, while reduce the impact of human factors, and increase the research on the impact of noun ambiguity on the visualization of nouns, so as to achieve the integrity of noun visualization in text visualization. In the follow-up work, we will study the visualization of nouns from sentence, paragraph and larger range.

References
[1] B Coyne and R Sproat. WordsEye: An Automatic text-to-scene conversion system. Proceedings of the SIGGRAPH 2001 Annual Conference on Computer Graphics, Los Angeles, CA USA, 2001, pp: 487-496.
[2] Zeng X, Mehdi Q H, Gough N E. From Visual Semantic Parameterization to Graphic Visualization. International Conference on Information Visualisation, IEEE Computer Society, 2005, pp: 488-493.
[3] F Yang, Y Zhou, X Luo. Scene layout in text-to-scene conversion. Proceedings of the 2nd International Conference on Systems & Informatics. United States, IEEE, 2015, pp: 891-895.
[4] Clay S R, Wilhelms J. Put: language-based interactive manipulation of objects. IEEE Computer Graphics and Applications, 1996, pp: 0-39.
[5] Sproat R. Inferring the environment in a text-to-scene conversion system. International Conference on Knowledge Capture, ACM, 2001, pp: 147-154.
[6] Chang A, Monroe W, Savva M, Christopher Potts, Christopher D. Text to 3D Scene Generation with Rich Lexical Grounding. Computer Science, 2015.
[7] Lu Ruqian, Zhang Songmao. From Story to Animation--Full Life Cycle Computer Aided Animation Generation. Acta Automatica Sinica, 2002, PP: 321-348.
[8] Huang Jie. The Entity Attribute Structure in Text-to-Scene Conversion. Chongqing University of Posts and telecommunications. 2018.
[9] Wu Jinbao. K-means Overview of algorithm research. Electronic technology & Software engineering, 2014, PP: 207-207.