STUDIA INFORMATICA 2010
Volume 31 Number 2A (89)

Tomasz KAPŁON
Politechnika Wrocławska, Instytut Informatyki, Automatyki i Robotyki

GRAFICZNA INTERPRETACJA ZNACZENIA TEKSTU

Streszczenie. Zaprezentowany został opis systemu typu text-to-picture generującego obraz w oparciu o tekst formułowany w języku naturalnym. System wykorzystuje pewne metody AI, w tym przetwarzanie języka naturalnego i soft computing oraz elementy grafiki komputerowej. Efektywność systemu sprawdzona została na podstawie dwóch scenariuszy opisu prostych scen. Rezultaty pokazały, że generowanie scen w oparciu o opis słowny daje szanse na zwiększenie możliwości komunikacji człowiek-komputer oraz człowiek-człowiek przy wykorzystaniu takich aplikacji np. w edukacji.

Słowa kluczowe: przetwarzanie języka naturalnego, przetwarzanie tekstu w obraz, reprezentacja wiedzy

GRAPHICAL INTERPRETATION OF TEXT MEANING

Summary. The text-to-picture system that synthesizes a picture from semi-general, (partially restricted) natural language text was presented. System using some AI methods, including natural language processing and soft computing and computer graphics. The effectiveness of the system was tested based on two simple scenarios. These results suggest that text-to-picture synthesis has potential in augmenting human-computer and human-human communication modalities, with applications e.g. in education.

Keywords: natural language processing, text-to-picture synthesis, knowledge representation

1. Wstęp

Obraz to więcej niż tysiąc słów. Niezależnie od języka komunikacji przekaz w postaci obrazu może być – przy pewnych założeniach, np. wspólne doświadczenia, znajomość tema-

tyki, dziedziny, problemu i związanym z nim pewnym aparatem pojęciowym – możliwy. Można by go uznać za uniwersalny, gdyby nie problem z obrazowaniem kontekstu czy różnym rozumieniem szczegółów, a w dużej mierze również z problemem przekazywania informacji na temat pojęć abstrakcyjnych – tu problem kontekstu, odczuwania i rozumienia jest szczególnie widoczny. Niemniej jednak zachodzi konieczność – zwłaszcza w świecie tak mobilnym, jak dzisiejszy – jednoznacznego przekazywania informacji szerokiej rzeszy ludzi, dla której nie istnieje jeden język komunikacji. Lotniska, drogi, uczenie języków, kontakt z ludźmi niepełnosprawnymi, szybka komunikacja w stanach zagrożenia, rozrywka, to dziedziny, w których pomocne, jeśli nie niezbędne do skutecznego działania, może być istnienie mechanizmów czy systemów komunikacji obrazowej, której efekt – jednoznaczny przekaz wizualny – powstaje na bieżąco – treść komunikatu nie jest znana w momencie zaistnienia potrzeby prezentacji – a co za tym idzie, musi być konstruowany na podstawie opisu słownego, w sytuacji idealnej, niezależnego od użytego języka (naturalnego), nie ograniczonego do słów kluczowych czy ustalonego zestawu pojęć i sformalizowanego sposobu ich zapisu. Czy takie systemy istnieją?

2. Systemy typu text-to-scene

NALIG [3], SPRINT [6], Put czy WordEye [4] to przykłady realizacji zadań klasy „text-to-scene" ograniczone w swoim działaniu do generowania scen w oparciu o szczegółowy opis sceny. Przykładami systemów typu „text-to-scene" są proponowane przez Lu [5] czy Mihalcea i Leong [2] czy Joshi, korzystające z zestawu słów kluczowych, jednak operujące na tekstach formułowanych w języku naturalnym. Te i podobne rozwiązania, np. SymWriter czy Blissymbols nie rozwiązują jednak problemu generowania scen w oparciu o dowolny tekst (*general text*). Bliższe temu są prace Goldberga [1]. Ich rozwiązanie ma u podstaw zdecydowanie silne rozwiązanie kwestii uczenia się i pozyskiwania wiedzy, jednak scena jest zbiorem obrazów, które nie tworzą spójnego wizualnie obrazu. Problem tkwi w braku generowania sceny spójnej wizualnie, czyli takiej, w której poszczególne obiekty (fragmenty obrazu) zachowują względem siebie związki, które znane są odbiorcy z natury.

Spójność wizualną rozumie się jako takie wzajemne ułożenie obiektów na scenie, które zachowuje pomiędzy nimi właściwe proporcje, takie, które odzwierciedlają ich wzajemne relacje w świecie rzeczywistym – istotne jest zachowanie proporcji, wielkości obiektów rzeczywistych, perspektywy, realności odzwierciedlenia.

3. Specyfika zadania i założenia

Prezentacja w formie obrazów znaczenia treści przekazu (np. tekstu) w postaci zrozumiałej dla większości odbiorców to nie tylko przedstawienie zestawu obrazków zaczerpniętych z bazy obrazów czy Internetu. Mając do dyspozycji pewien model reprezentacji wiedzy w postaci zdań języka nietrudno jest wykonać właściwą i poprawną analizę treści (zarówno syntaktyczną, jak i semantyczną), a następnie wyświetlić kolejne obrazy reprezentujące role semantyczne pełnione w zdaniu przez kolejne grupy wyrazowe. Dobrze byłoby, aby obraz (scena bądź animacja) stanowiła wizualną całość, aby istniała zbieżność wyglądu czy skali poszczególnych fragmentów, aby obraz, który zostanie wygenerowany, przekazywał informacje w sposób jednoznaczny, a nie stanowił zagadki samej w sobie. Podstawowe rozwiązanie powinno dostarczać obrazów jednoznacznie wobec przekazu (w postaci opisu słownego), którego treść pozbawiona jest elementów abstrakcyjnych, dla których znalezienie materialnych odpowiedników nie jest możliwe. Rozwiązanie powinno kreować scenę złożoną z obiektów podstawowych (takich, które da się opisać wzorem, np. podstawowe figury geometryczne). Zaletą tychże jest łatwość zapisu w bazie wiedzy, łatwość modyfikacji (skalowanie, przesunięcie, obrót) oraz nieskomplikowany sposób tworzenia kształtów złożonych reprezentujących obiekty materialne.

3.1. Założenia

Reasumując, zadanie polega na wygenerowaniu sceny w oparciu o opis sformułowany w postaci zdań języka naturalnego i scena ta ma być spójna wizualnie. System w wersji podstawowej posiada możliwość analizy syntaktycznej i semantycznej tekstu, wygenerowania obrazów semantycznych zdań tekstu oraz znalezienie odpowiedników obrazów semantycznych w bazie zawierającej opisy obiektów. Liczba obiektów ograniczona jest do podstawowych figur geometrycznych z podstawowymi parametrami charakterystycznymi.

4. Proponowane rozwiązanie

Prezentowany system posiada elementy wspólne z wymienionymi w punkcie 2. Przede wszystkim do tworzenia scen (animacji) wykorzystywane będą opisy kreowania obiektów i być może, w przypadkach braku możliwości lub trudności w opisie, modele obiektów i/lub grafiki same w sobie. W kolejnych rozwinięciach implementowane będą mechanizmy uczenia się i pozyskiwania wiedzy ze źródeł zewnętrznych (Internet). Niemniej jednak najważniejsza pozostanie kwestia generowania sceny/animacji spójnej wizualnie. Niezbędne będzie

również, co wykracza poza ramy tego opracowania, zweryfikowanie lub zmiana modelu reprezentacji wiedzy. W przypadku omawianego systemu analiza syntaktyczna odbywa się w oparciu o model formalny Zmodyfikowanej Gramatyki Łączeń [9] oraz o model formalny Obrazowej Reprezentacji Semantyki Zdań [7, 8]. Scenariusz składający się z dowolnej liczby zdań opisuje scenę. Rozważmy zdanie.

T*he Gray block is set on a ground*. (1)

Podział na grupy słów zdania (1) jest następujący (rdzeń grupy wytłuszczony):

the gray ***block*** – $d_0\ d_1\ d_2 \leftrightarrow a_1$

is *set* – $d_3\ d_4 \leftrightarrow a_2$

on a ***ground*** – $d_5\ d_6\ d_7 \leftrightarrow a_3$

$s_n = a_1\ a_2\ a_3$,

gdzie d_j reprezentują wyrazy, a_i grupy słów i s_n zdanie.

Poszczególne grupy i zdania gromadzone są w bazie wiedzy w postaci wyrażeń symbolicznych. Dla zdania (1) wyrażenie ma postać:

$$A^+ = {}^0(\ a_1\ {}^1(\ z_1 s_1 a_2\ {}^2(\ z_1 s_1 a_3\ {}^3(\ z_3 s_1 a_1\ )^3\ )^2\ )^1\ )^0$$

Poprawność zdania (1) weryfikowana jest w oparciu o zmodyfikowaną gramatykę łączeń. Aby zdanie można było uznać za poprawne, wystąpić muszą dwa warunki konieczne, zdanie musi dać się podzielić na grupy słów[1] oraz muszą istnieć odpowiednie – wcześniej zdefiniowane dla danego języka – połączenia pomiędzy odpowiednimi grupami. Więcej na temat algorytmu weryfikacji można znaleźć w pracy [9].

Analiza semantyczna zdania polega na nadaniu rdzeniowi grupy cechy semantycznej i na jej podstawie określenia kategorii semantycznej grupy [8].

the gray ***block*** $\leftrightarrow a_1$ – AGENT

is *set on* $\leftrightarrow a_2$ – ACTION

a ***plane*** $\leftrightarrow a_3$ – LOCATION

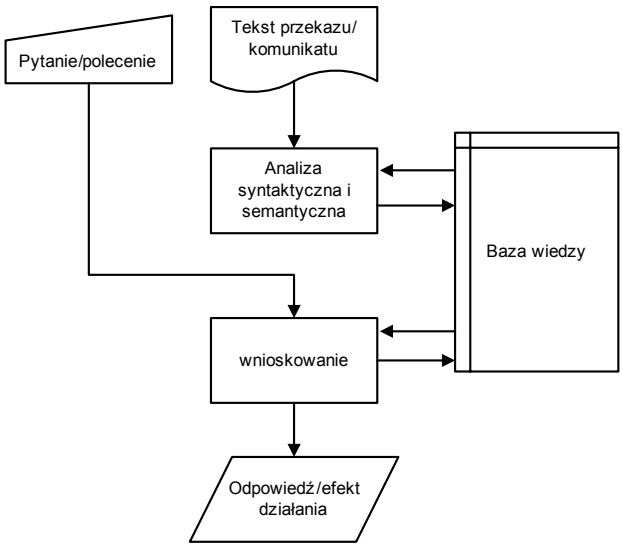
Następnie określa się semantykę c_m dla zdania, strukturę semantyczną zdania na poziomie cech semantycznych $B^+(c_m)$, następnie tworzy się wyrażenie $U^+(c_m)$ określające zależność pomiędzy strukturą syntaktyczną $A^+(c_m)$ i semantyczną $B^+(c_m)$.

W ten sposób w bazie wiedzy zgromadzona jest wiedza na temat struktury składniowej i znaczeniowej zdań tworzących scenariusz. Zdefiniowane w [8] operacje na tych strukturach pozwalają na wnioskowanie – generowanie odpowiedzi na pytanie w przypadku systemu wnioskującego – oraz, w połączeniu z częścią zawierającą opisy obiektów, na właściwą interpretację znaczenia (pełnionej w zdaniu roli) poszczególnych grup słów.

[1] Funkcjonalna grupa słów zawiera od 1 do *n* słów, które w zdaniu pełnią funkcję rzeczownikową, czasownikową itp.

5. Budowa i działanie systemu

System generowania scen będących graficzną reprezentacją znaczenia przekazu zawiera elementy tradycyjnego systemu przetwarzania treści formułowanej w języku naturalnym (rys. 1). Przekaz trafia do modułów analizy syntaktycznej i semantycznej, wyniki analiz gromadzone są w bazie wiedzy, w której przechowywane są również opisy tworzenia obiektów graficznych, stanowiące jedynie część faktycznego opisu obiektu. Jak wspomniano, nietrudno jest wyświetlić obiekt na scenie, trudność polega na jego właściwym skorelowaniu z innymi obiektami. Zadanie to realizowane jest w oparciu o szczególną analizę semantyczną, której celem jest odkrywanie istnienia i wpływu pojęć stanowiących o wyglądzie i wzajemnym oddziaływaniu obiektów na scenie (np. wielkość, położenie). Pojęcia te w zdecydowanej większości są pojęciami, jak na język naturalny przystało, rozmytymi, a co za tym idzie, opisującymi wielkości zależne od kontekstu, w którym występują. Problem reprezentacji w bazie opisów obiektów stanowi oddzielne zagadnienie, zdecydowanie wybiegające poza ramy niniejszego opracowania. Przyjęto zatem, że istnieją one w bazie wiedzy w formie ustalonej i pełnej.



Rys. 1. Ogólny schemat przetwarzania tekstu w obraz
Fig. 1. General scheme of text-to-scene processing

5.1. Reprezentacja obiektów

Dla każdego obiektu zdefiniowane zostały domyślne wartości parametrów charakteryzujących poszczególne obiekty: wysokość, szerokość, głębokość, kolor, pozycja, przezroczystość. Podstawowy opis obiektu zawiera: nazwę, pozycję punktu charakterystycznego (x_p, y_p), wymiary i kolor. Każdy obiekt w opisie ma właściwą dla niego (nadaną a priori) kategorię

semantyczną (AGENT, ACTION, …) i cechę semantyczną (*ANIMAL*, *FOOD*, *PROCESS*, …).

Proces wyboru (interpretacji znaczenia) odpowiednich obiektów następuje przez skojarzenie kategorii i cechy charakteryzującej pojęcie w zdaniu z odpowiednimi argumentami w opisie obiektu. Obiekty posiadają zapisane podstawowe parametry domyślne, a w przypadku rozpoznania nazwy parametru związanego z obiektem w zdaniu następuje, jeśli to konieczne, modyfikacja np. wartości argumentu (por. scenariusz 1: dwa pierwsze zdania). Obiekty w formie graficznej, zgodnej ze sposobem ich utworzenia (np. w formie wzoru na sferę), umieszczane są na scenie. Jeżeli w bazie nie istnieje szukany obiekt, na scenie umieszczany jest komunikat z nazwą pojęcia, którego odpowiednika w bazie nie ma (por. rysunek 3).

6. Wyniki

Badania wykonane zostały w oparciu o dwa scenariusze. Pierwszy testowy, obejmował tekst, który został przygotowany przez autora ze znajomością zawartości bazy wiedzy i możliwości generowania scen. Scenariusz posłużył do sprawdzenia poprawności generowania scen. Drugi scenariusz zawierał tekst przygotowany przez grupę testową – osoby znające ograniczenia dziedzinowe i przeznaczenie aplikacji. Zakres pojęciowy ograniczony został do dziedziny, klasycznego skądinąd przykładu, Blocks World – zestawu pojęć obejmującego figury geometryczne 3D. Wydaje się być prawdziwem, że rozszerzenie zakresu pojęciowego – w ramach, jak wspomniano we wstępie, obiektów posiadających swoje materialne odpowiedniki – nie spowoduje utraty zdolności systemu do poprawnego generowania scen.

Największym wyzwaniem pozostaje nadal interpretacja roli semantycznej ACTION. W zasadzie to na podstawie grup czasownikowych następuje kreacja sceny. Trudność interpretacji polega na poprawnym zdefiniowaniu w bazie wiedzy sposobu przedstawienia działania, które de facto nie ma fizycznego odpowiednika i konieczne jest stworzenie iluzji ruchu, kierunku wykonania akcji, jej aktorów i narzędzi (często niebędących obiektami materialnymi). W przypadku przygotowanych scenariuszy nie miało to znaczenia, jednak założeniem jest poprawna interpretacja i generowanie scen (animacji) niezależnie od treści, a przynajmniej uzyskania poprawności w zakresie scen (animacji) informacyjnych czy wspomagających procesy np. nauczania.

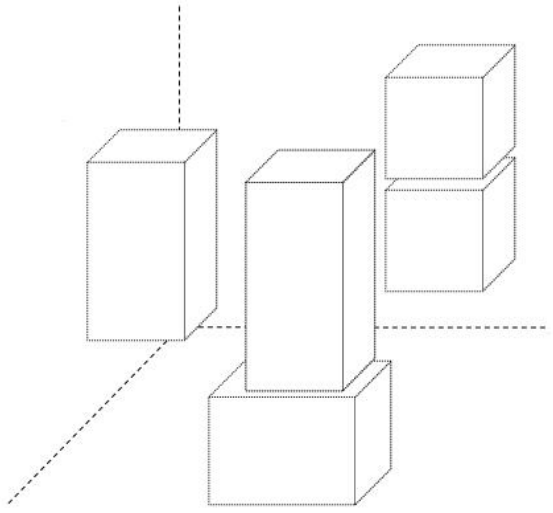
6.1. Scenariusz pierwszy

Pełna treść scenariusza pierwszego zawierała 12 zdań prostych opisujące scenę i zawierała precyzyjny opis sceny (rozmieszczenie obiektów).

Kostka ma 2 centymetry szerokości,2 centymetry wysokości i 2 centymetry długości. Na płaszczyźnie leży kostka o szerokości 3,2 centymetra i wysokości 2 centymetrów. Na niej stoi kostka wyższa od 2 centymetry. 3 centymetry za nią stoi kolejna kostka. Nad nią, na wysokości 42 milimetrów nad płaszczyzną wisi kostka. Blok ma 2 centymetry szerokości i 4 centymetry wysokości. Pozycja bloku to (0, 4, 2). Blok wisi.

W scenariuszu zawartych jest 7 zdań złożonych, które w procesie analizy rozkładane są (i tak zapamiętywane w bazie wiedzy) na zdania proste. Scenariusz nie obejmuje użycia pojęć rozmytych, a przez to nie jest dokonywana modyfikacja w oparciu o ich interpretację. Jest to zabieg celowy, ponieważ jest to zadanie samo w sobie szczególnie rozbudowane i ze swojej natury mocno zależne od kontekstu. Użycie pojęć rozmytych spowodowałoby trudność w sprawdzeniu poprawności działania systemu.

Efektem działania była scena widoczna na rysunku 2. Ponieważ był to scenariusz z danymi spreparowanymi, wynik działania był prawidłowy. Ponieważ opisy w bazie wiedzy były dwujęzyczne, więc treść scenariusza wprowadzono w języku polskim, pomimo tego, że algorytmy analizy syntaktycznej i semantycznej stworzone zostały dla języka angielskiego. Dodatkowo, treść przekazu zawierała konkretne miary, które jednoznacznie opisywały wygląd sceny. Położenia domyślne dla kostki i bloku są ustalone w opisie obiektu.

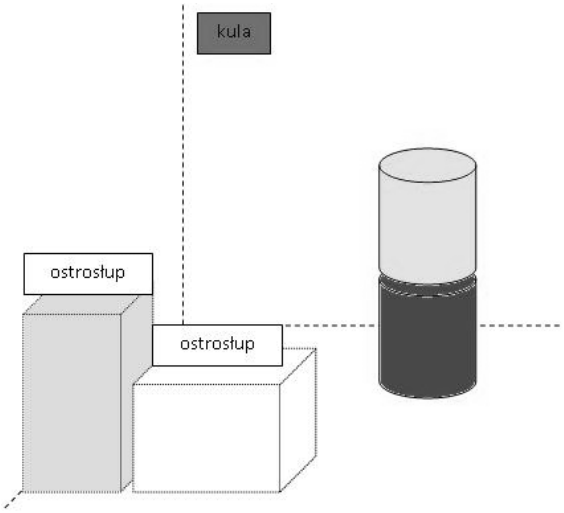


Rys. 2. Efekt interpretacji scenariusza pierwszego
Fig. 2. First scenario visualization

6.2. Scenariusz drugi

Treść scenariusza drugiego zawierała 18 zdań prostych. Ponieważ część bazy, w której gromadzona jest wiedza o składni i obrazach semantycznych tekstów, nie posiada ograniczeń związanych z opisem obiektów – została stworzona wcześniej i ma już ustaloną formę – można wprowadzić i analizować praktycznie dowolny tekst. Część obrazowa jest nie w pełni ukształtowana i nie posiada wielu elementów – nawet z zakresu, który został określony jako dopuszczalny, czyli zestawu obiektów geometrycznych 3D – stąd też scena opisana w scenariuszu nie została do końca poprawnie wygenerowana (rys.3), brak obiektów zastąpionych etykietami.

Grey block is set on the ground. The ground is white. On the right is a white block. The sphere has a radius of 3cm and has a yellow color. Pyramid stands at a height of 3cm above the plane and is purple. On the right side there is another pyramid. The sphere is blue and hangs over the pyramid. Cylinder is red and stands on the ground. Cylinder is yellow. Another cylinder stands above the cylinder. A monkey is huge and ugly. The monkey eats a banana and has a hat on its head.



Rys. 3. Efekt interpretacji scenariusza drugiego
Fig. 3. Second scenario visualization

7. Zakończenie

Zadanie interpretacji znaczenia treści i na tej podstawie generowania spójnego wizualnie obrazu zostało – przynajmniej w ogólnym zarysie – przedstawione w niniejszej pracy. Elementem, który można wskazać jako nowy w rozwiązywaniu zadania poprawnej prezentacji znaczenia treści przekazu jest zdolność generowania scen spójnych wizualnie. Sformułowano podstawowe założenia dotyczące budowy i działania systemu. Przedstawiono wyniki działa-

nia w oparciu o dwa proste scenariusze opisu scen. Pokazano możliwości generowania scen spójnych wizualnie. Kolejne prace skupiać się będą na rozszerzeniu możliwości generowania scen i animacji, zwiększeniu liczby opisanych obiektów, sprawdzeniu zaproponowanego modelu na tekstach zawierających pojęcia rozmyte, jak i na tekstach rzeczywistych oraz weryfikacji efektów działania przez użytkowników testowych.

Podstawową zaletą rozwiązania prezentowanego jest fakt operowania na obiektach zdefiniowanych w bazie wiedzy, które są, przed wyświetleniem, modyfikowane w oparciu o fakty zawarte (o ile są) w tekście. Rozwiązania znane z literatury operują na obrazach lub elementach obrazów rastrowych często poszukiwanych ad hoc w Internecie, co nie pozwala na utrzymanie postulowanej w tym rozwiązaniu spójności wizualnej. Porównanie działania i ocena jakości – zapewne subiektywna – z rozwiązaniami spotykanymi w literaturze jest na obecnym etapie trudna. Powodem jest uboga baza wiedzy w części obrazowej. Niedostatek modeli – widoczny chociażby w scenie wygenerowanej dla drugiego scenariusza – powoduje, że nie da się pokazać w pełni możliwości (na obecnym etapie mimo wszystko koncepcyjnych), a w związku z tym porównanie nie byłoby obiektywne.

BIBLIOGRAFIA

1. Goldberg A.B., Xiaojin Z., Dyer C. R., Eldway M., Heng L.: Easy as ABC? Facilitating Pictorial Communication via Semantically Enhanced Layout. In Twelfth Conference on Computational Natural Language Learning (CoNLL), 2008.
2. Mihalcea R., Leong B.: Toward Communication Simple Sentences Using Pictorial Representations. In Proc. Conf Association for Machine Translation in the Americas (AMTA), 2006.
3. Adorni G., Manzo M. D., Giunchiglia F.: Natural language driver image generation. In Proc. COLING, 1984.
4. Coyne B. Sproat R.: WordsEye: An automatic text-to-scene conversion system. In Proc. SIGGRAPH, 2001.
5. Lu R., Zhang S.: Automatic Generation of Computer Animation: Using AI for Movie Animation. Lecture Notes in AI, Berlin: Springer-Verlag, Vol. 2160, 2001.
6. Yamada A., Yamamoto T., Ikeda H., Nishida T., Doshita S.: Reconstructing spatial image from natural language texts. In Proc. COLING, Vol. 4, 1992.
7. Kapłon T.: Model formalny obrazowej reprezentacji semantyki zdań języka naturalnego do wnioskowania przez komputer. Raport seria PRE 30/2003 (praca doktorska), Oficyna Wydawnicza Pol. Wrocławskiej, 2003.

8. Kapłon T.: Model formalny obrazowej reprezentacji semantyki zdań języka naturalnego. Bazy danych. Rozwój metod i technologii. Bezpieczeństwo, wybrane technologie i zastosowania.Praca zbiorowa pod red. Stanisława Kozielskiego., WKŁ, Warszawa 2008.
9. Mierzwa J.: Model formalny komputerowej dekompozycji zdań języka naturalnego na grupy słów do celów wnioskowania przez komputer. Raport seria PRE 39/2001 (praca doktorska), Oficyna Wydawnicza Pol. Wrocławskiej, 2001.

Recenzenci: Dr inż. Małgorzata Bach
Dr inż. Michał Kawulok

Wpłynęło do Redakcji 31 stycznia 2010 r.

Abstract

Text-to-picture systems are created out of a dedicated systems limited the range of action (Liu, Mihalcea) or learning systems, however, with a poorly developed mechanism of visualization (Goldberg). The text-to-picture system which generates a scene or animation on the basis of the text formulated in natural language was presented. The system uses existing models of the formal representation of semantics of sentences and modified link grammar. The very process of text processing is described in [7, 8, 9], and the result of this process is semantic image, which has its counterpart in the part of the knowledge base containing the descriptions of graphical objects. The objects create the scene (Figure 2 and 3). The algorithm of selection of objects to display on the stage associates semantic category and semantic features of concepts and desired object. Modifying the arguments that describes the objects is implemented in the interpretation of the content. Potential modifying arguments are associated with objects. If there is a correlation, object parameters are changed. Efficiency of the system was tested based on two scenarios describe simple scenes. The results showed that the generation of scenes based on a verbal description is possible and gives a chance to increase the opportunities for communication and human-computer human-human.

Adres

Tomasz KAPŁON: Politechnika Wrocławska, Instytut Informatyki, Automatyki i Robotyki, Wybrzeże Wyspiańskiego27, 50-370 Wrocław, Polska, tomasz.kaplon@pwr.wroc.pl.