On 'Speaking-Cameras'

A. Fexa[1], K. Uchimoto[2], H. Isahara[2], J. Kawai[2], and H.-H. Nagel[1]

[1] Institut für Algorithmen und Kognitive Systeme, Fakultät für Informatik, Universität Karlsruhe (TH), 76128 Karlsruhe, Germany
{fexa|nagel}@iaks.uni-karlsruhe.de
[2] National Institute of Information and Communications Technology (NICT), Kyoto, Japan
{uchimoto|isahara|jkawai}@nict.go.jp

The notion of a 'Speaking-Camera' refers to an experimental system which transforms the movement of road vehicles recorded by a stationary video-camera into natural language textual descriptions [7, 6]. Here, we emphasize system *extensions* to incorporate, first, text generation for the *Czech* language [2], followed by the addition of a capability to generate texts in the *Japanese* language [3]. Figure 1 shows a coarse system outline, details can be found in [3].

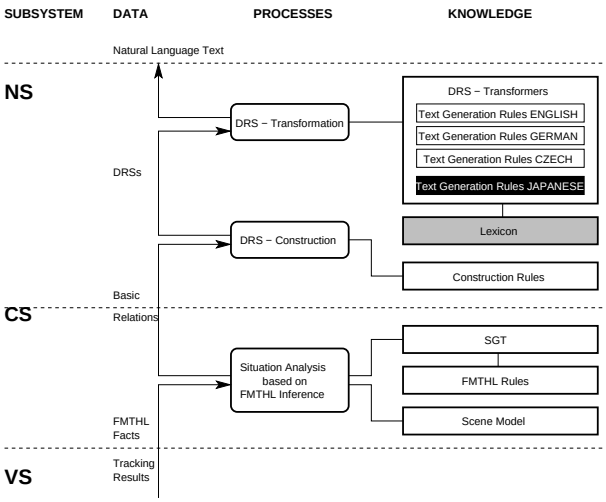


Fig. 1. Extension of the Natural Language Subsystem (NS) in order to generate, too, natural language descriptions of traffic developments in the *Czech* and *Japanese* Language (adapted from [5, 2]). The 'Conceptual Subsystem (CS)' comprises a schematic representation of system capabilities, of the environment, and of admissible spatio-temporal changes, in particular a representation of the behavior of agents acting within the recorded scene. These schematic representations are based on a Fuzzy Metric-Temporal Horn Logic (FMTHL), an extension of First Order Predicate Logic by fuzzy predicates and metric-temporal attributes, see [8]. A Computer Vision Subsystem (VS) – see [3] – provides stepwise transformations between video-signals recorded from the system environment and conceptual representations at the CS.

Results obtained for the `'dtneu05'` video sequence are illustrated by Figure 2 (left panel). The thick black line indicates the trajectory of the vehicle whose movements are described by the automatically generated text. Analogous results for the `'tankstelle'` sequence are shown in the right panel of this same Figure 2. A sketch of the vehicle trajectory has been overlaid to a sketch of (parts of) the gas-station premise in order to visualize the vehicle trajectory during the subsequence covered by the text generation. Czech Language Text Generation (CLTG) as well as Japanese Language Text Generation (JLTG) are appropriate regarding the content and are correct with respect to the linguistic aspect.

So far, modifications necessary in order to implement CLTG and JLTG could be restricted to components located in the NS, specifically in the lexicalisation rules, the morphology, the Text-Generation rules, and the orthography, in addition to an extension of the lexicon by appropriate entries. The components that had to be changed for JLTG include components which have been language specific already for English, German, and Czech. In addition to these already language-specific components, a component for orthography had to be changed during the implementation of JLTG. Note that *it has not been necessary yet to modify system components related to the conversion of signals to conceptual representations or related to the treatment of vehicle behavior.*

The essentially rule-based approach chosen for the original system designed for German Language Text Generation (GLTG) [4] turned out to offer sufficient flexibility to facilitate adaptation of this system *without great problems to three additional languages, two of which are substantially different from German.*

An alternative to an implementation of language-specific NLG-components as pursued here could have been the generation of textual descriptions from videos *in only one language*. The resulting text could then be converted by Machine Translation (MT) systems into a great variety of other languages. This latter approach implies that the developments extracted from videos would have to be converted into some kind of more or less explicit intermediate representation *specific to each MT-system*. In our approach, this system-internal representation (see CS in Figure 1) can be exploited not only for the Natural Language Generation (NLG)-components, but at the same time serves for feedback into the quantitative image sequence evaluation processes [1].

References

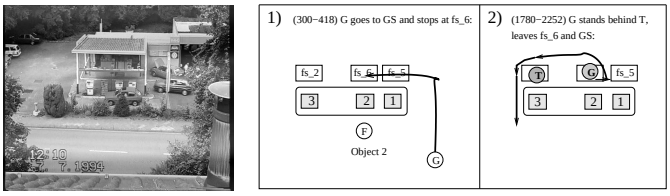
1. M. Arens and H.-H. Nagel. Quantitative Movement Prediction Based on Qualitative Knowledge about Behavior. *KI Künstliche Intelligenz*, 2/05:5–11, May 2005.
2. A. Fexa. Dependence of Conceptual Representations for Temporal Developments in Videosequences on a Target Language. In D. Paulus and D. Droege, editors, *KI 2005 Workshop 7 'Mixed-reality as a Challenge to Image Understanding and Artificial Intelligence'*, pages 47–54, Koblenz, Germany, 11 September 2005. Fachberichte Informatik, Universität Koblenz-Landau, Institut für Informatik, Universitätsstr. 1, 56070 Koblenz, Germany. ISSN 1860-4471.
3. A. Fexa, K. Uchimoto, H. Isahara, J. Kawai, and H.-H. Nagel. Speaking Cameras. Technical report, Institut für Algorithmen und Kognitive Systeme, Universität Karlsruhe (TH), 76128 Karlsruhe, Germany, March 2006.

The white car comes in from the Bernhard Street. It follows another car. It crosses the intersection. It turns left into the Chapel Street.

Das weisse Fahrzeug kommt aus der Bernhardstrasse. Es folgt einem anderen Fahrzeug. Es ueberquert die Kreuzung. Es faehrt links in die Kapellenstrasse.

`Bílé auto přijíždí z Bernhardstrasse. Jede za jiným autem. Přejíždí přes křižovatku. Zahýbá doleva do Kapellen-strasse.`

白い車がBernhard通りから来ます。
その車は他の車に続きます。
その車は交差点にさしかかります。
その車はChapel通りの方に左折します。



The car drives to the first pump. Now it has reached the first pump. It moves on to the second pump. Now it has reached the second pump. It stops. Now it drives off to the exit. Therefore it has to overtake another car. It backs up. It stops. Now it drives towards the other car. It overtakes the other car. It leaves the other car. It drives to the exit. Now it has reached the exit.

Das Fahrzeug faehrt zu der ersten Zapfsaeule. Jetzt hat es die erste Zapfsaeule erreicht. Es faehrt weiter zu der zweiten Zapfsaeule. Jetzt hat es die zweite Zapfsaeule erreicht. Es haelt an. Jetzt faehrt es zu der Ausfahrt weiter. Dazu muss es ein anderes Fahrzeug ueberholen. Es setzt zurueck. Es haelt an. Jetzt faehrt es auf das andere Fahrzeug zu. Es ueberholt das andere Fahrzeug. Es verlaesst das andere Fahrzeug. Es faehrt zu der Ausfahrt. Jetzt hat es die Ausfahrt erreicht.

`Auto jede k první pumpě. Dojelo k ní. Jede dále k druhé pumpě. Dojelo k ní. Zastavuje. Jede k výjezdu. Kvůli tomu musí předjet jiné auto. Couvá. Zastavuje. Jede k tomu autu. Předjíždí ho. Odjíždí od něj. Jede k výjezdu. Dojelo k výjezdu.`

車が初めの給油器の方に向かいます。 今その車は初めの給油器に到着しました。
その車は二番目の給油器に移動します。 今その車は二番目の給油器に到着しました。
その車は止まります。 今その車は出口に向かいます。
そのためにはその車は他の車を追い越さなければなりません。
その車はバックします。 その車は止まります。 今その車はその別の車の方に向かいます。
その車はその別の車を追い越します。 その車はその別の車から離れます。
その車は出口の方に向かいます。 今その車は出口に到着しました。

Fig. 2. Two examples of text generated for the `'dtneu05'` (left column) and the `'tankstelle'` video sequence (right column with a screen shot and a schema of movement for car 'G'). The images have been taken from [5, 4] and modified slightly.

4. R. Gerber. *Neustrukturierung der Generierung von Text aus Ergebnissen der Bildfolgenauswertung*. Interner Bericht (in German), Institut für Algorithmen und Kognitive Systeme, Fakultät für Informatik der Universität Karlsruhe (TH), 76128 Karlsruhe, Germany, September 2004.
5. R. Gerber. On Switching the Discourse Domain for Text Generation from Videos. *Cognitive Vision System - Final Report (Draft of 30 November 2004)*, pages 347–362, November 2004.
6. R. Gerber and H.-H. Nagel. Discourse Representation Theory for Generating Text from Video Input. Technical report, Institut für Algorithmen und Kognitive Systeme, Universität Karlsruhe (TH), 76128 Karlsruhe, Germany, January 2005.
7. H.-H. Nagel. Steps toward a Cognitive Vision System. *AI-Magazine*, 25(2):31–50, Summer 2004.
8. K.H. Schäfer. *Unscharfe zeitlogische Modellierung von Situationen und Handlungen in Bildfolgenauswertung und Robotik*, volume 135 of *Dissertationen zur Künstlichen Intelligenz (DISKI)*. infix–Verlag, Sankt Augustin, Germany, 1996. (Dissertation, Fakultät für Informatik, Universität Karlsruhe (TH), Juli 1996).