Hybrid Natural Language Generation for Data2Text in Portuguese

José Casimiro Pereira[1], António Teixeira[2], Joaquim Sousa Pinto[2]

[1]IEETA, University of Aveiro, [2] Department of Electronics, Telecommunications and Informatics/IEETA, University of Aveiro

Abstract

In many new interactions with machines, such as dialogue or output using voice, there is the need to convert information internal to a system into sentences. Natural Language Generation Systems (NLG) is a possible solution to achieve this conversion.

A Data2Text systems are a sub-type of NLG systems suitable when input is nonlinguistic data. Although there are some successful experiences in data2text, fast development of such systems is difficult and usually demanding in terms of resources, knowledge and development time.

Our main objective is to make possible to create data2text systems for real applications with a minimum of knowledge and resources and in a reduced time, without compromising quality and variety of the generated sentences.

With this objective in mind, we proposed and have been developing a hybrid data2text, combining translation based NLG module, a template-based NLG module and a sentence quality evaluation module capable of providing information on the Intelligibility or Quality of the sentences.

The proposal has been applied with good results to a Medication Assistant for Smartphones and is now being applied to a Tourism scenario (Hotels evaluation).

Data2Text

In many new interactions with machines, such as dialogue or output using voice, there is the need to convert information internal to a system to sentences. An example is shown in Fig. 1. This conversion can be done using Natural Language Generation (NLG) or a subtype of those systems, the Data2Text systems, designed to have data as input.

The Challenges

Creation of new Data2Text systems by developers is not an easy task due to: (1) lack of tools to support in the development; (2) the scarcity of resources for most languages and domains; (3) the large amount of knowledge that classic approaches demand, particularly in Linguistics.

On top of these problems, classical approaches, such as template-based methods, fail to provide good output variability, a requirement for high quality interaction with users, and development time is too long to be usable in many applications.

Our proposal: Hybrid Natural Language Generation

Systems based on automatic translation provide sentences with the important variability, but this doesn't come without a cost. Contrary to template-based, these systems produce sentences with heterogeneous quality.

We proposed and explored the combination of a translation based NLG system with a classifier module capable of providing information on the Intelligibility or Quality of the sentences. Sentences marked as unacceptable are replaced by template-based generated ones.

In Fig. 2 is presented the architecture of our system. It is based in two main modules, Translation module and Template module, assisted by a classifier based sentence quality estimation module, combining extraction of linguistic features with a classifier trained in a manually annotated corpus.

The Translation Based NLG Module uses a machine translation system (Moses) to translate from a vector with system internal data to a sentence in Portuguese. In Fig. 3 is outlined the phrase-base translation method.

When the quality estimation module mark the sentence produced by Moses as having low quality, a sentence produced by a Template Based NLG Module is used. This module was developed using XML and XSLT.

Results

We adopted for our development and evaluation, the scenario of providing assistance on taking medication using a smartphone (Fig. 1).

Evaluation of the translation based module revealed most of the sentences as Intelligible (Fig. 4). Low quality sentences are less than 30%. Human evaluators used more often the scores between 3 and 5 (in a scale 1 to 5), evaluating as Good or Excellent 46 % of the sentences and as Excellent 25% of them. Examples of produced sentences are presented in Fig. 5.

Results regarding evaluation of the sentence quality estimation module suggest that our approach is valid, as best results obtained have false positives below 8% and can be lower in practical application (around 3%).

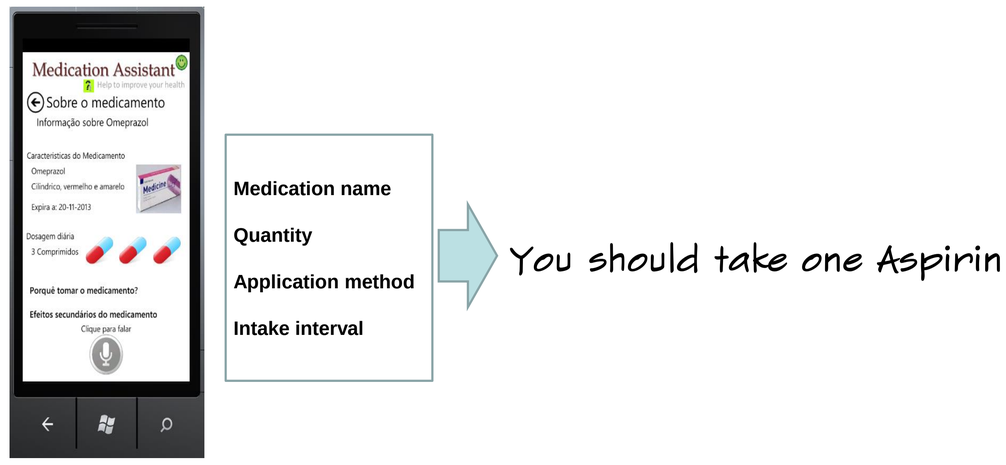

Fig 1 / Application example.

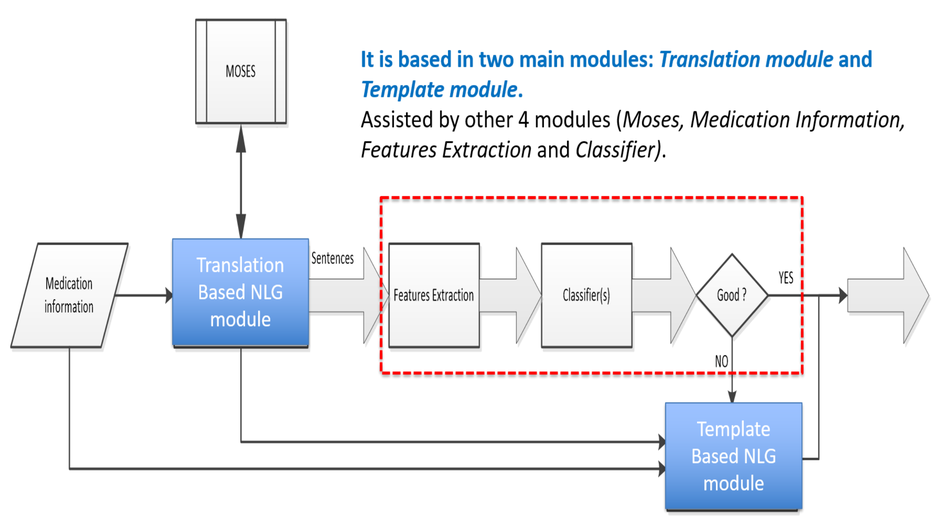

Fig 2 / Hybrid Natural Language Generation System.

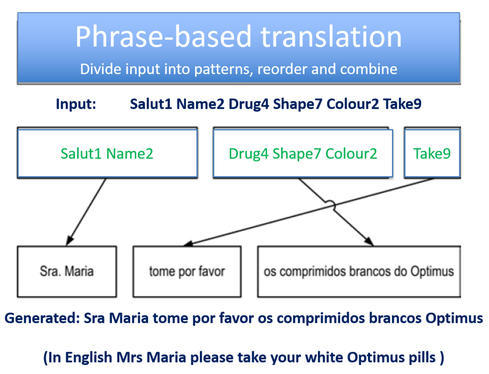

Fig 3 / An example of phrase-based translation. The example is taken from our Medication Assistant application scenario.

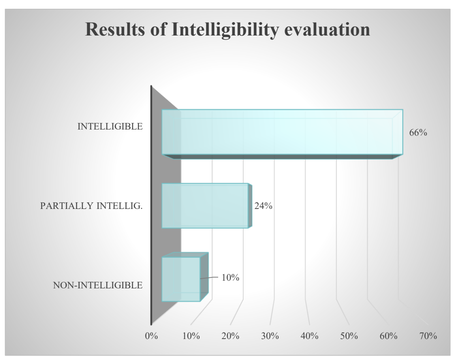

Fig. 4 / Results from intelligibility evaluation.

Senhor Paulo tome dois comprimidos de Salazopirina
(input: person61… medication4 … quantity2…)

Marcelo aplique ao tomar a cápsula branca e azul de dez horas em dez horas

Augusto ao acordar aplique a bomba de inalação pulmicort

Fig. 5 / Examples of sentences generated by the translation based module (in Portuguese).

Conclusion

With the proposed system, the limitations regarding uniformity of quality of the translation based generation are addressed while preserving the required variability of the sentences produced. The development of the translation based module does not demand more than basic knowledge in areas such as Linguistics. As the quality of the sentences is in general good, there is no need for the development of a very complete, and complex, template-based generation module.

Recent Publications

- José Casimiro Pereira, António Teixeira. Geração de Linguagem Natural para Conversão de Dados em Texto - Aplicação a um Assistente de Medicação para o português. LinguaMática, July 2015.
- José Casimiro Pereira, Joaquim Sousa Pinto, António Teixeira. Towards a Hybrid NLG System for Data2Text in Portuguese. CISTI 2015, June 2015.
- António Teixeira, José Casimiro Pereira, Pedro Francisco, Nuno Almeida. Tradução Automática na Interação com Máquinas. Linguística, Informática e Tradução: Mundos que se Cruzam - Homenagem a Belinda Maia, University of Oslo, Oslo, Norway, 2015
- António Teixeira, Flávio Ferreira, Nuno Almeida, Samuel Silva, Ana Filipa Rosa, José Casimiro Pereira, Diogo Vieira. Design and Development of Medication Assistant Elderly-centred Design to Go Beyond Simple Medication Reminders. Universal Access in the Information Society, 2016.

Partial funding from European Union's Seventh Framework Programme for research, technological development and demonstration, under grant agreement nº 610986 (project IRIS, FP7-PEOPLE-2013-IAPP)
http://www.iris-interaction.eu/.

Work also partially funded by National Funds through FCT - Foundation for Science and Technology, in the context of the projects UID/CEC/00127/2013.