PhD thesis: "Hierarchical language models based on classes of phrases: formulation, learning and decoding". (Original: "Modelos de lenguaje jerárquicos basados en clases de *phrases*: formulación, aprendizaje y decodificación.")

Author: *Raquel Justo.* **Supervisor:** *M. Inés Torres*

Department of Electricity and Electronics. University of the Basque Country. Spain
raquel.justo@ehu.es, manes.torres@ehu.es

1. Commitee

- Renato de Mori. Professor of University of Avignon (France).
- Jos Miguel Bened Ruiz. Professor Technical University of Valencia.
- Eduardo Lleida Solano. Professor University of Zaragoza.
- Emilio Sanchís Arnal. Technical University of Valencia.
- Javier Ferreiros Lopez. Technical University of Madrid.

Qualification: with honors (cum laude)

2. Abstract

This thesis focuses on the area of stochastic language modeling. A stochastic language model captures the way in which the combination of words is carried out in a specific language. It does so by making use of probability distributions of linguistic events, such as the frequency of appearance of words in sentences. Large amounts of training data, not always available, are required to get a robust estimation of the parameters defining such models.

In this work, a two-level hierarchical language model, based on classes of phrases, is proposed to deal with data sparseness. Each level in the model is associated to a different knowledge source. In the upper level the relations among classes are taken into account, i.e. relations among abstract entities employed to generalize. In the second level the relations among words are considered. The cooperation between different levels allows to build an improved language model. Within this framework different approaches and ways of combining models are defined and formulated.

Through out this work language modeling has been explored in the framework of Automatic Speech Recognition (ASR). Thus, a methodology to integrate the proposed models into the decoding stage of the ASR system has been developed.

In order to validate the presented approaches an experimental stage has been carried out using different databases. Three different languages and tasks of different complexity, spontaneous speech and read speech,... have been employed.

On the other hand, the use of the proposed hierarchical language models within a dialogue system prototype has been explored. In this case the main goal is to maximize the performance of the system in real working conditions.

Finally, a translation model based on the same hierarchical nature has been defined and formulated. This model has been integrated into a speech translation system. The methodology employed to integrate the language model in the ASR system can be directly applied to this case.

3. Curriculum Vitae

1. PERSONAL DETAILS

   Raquel Justo Blanco
   Departament of Electricity and Electronics
   University of the Basque Country.
   48940 Leioa. Spain.
   +34 946015364
   raquel.justo@ehu.es

2. EDUCATION:
   - Electronics Engineer Degree from the University of the Basque Country in 2001
   - Bachelor's degree in Physics from the University of Cantabria in 2004.
   - PhD degree in Language and Computation Systems from the University of the Basque Country in 2009.
3. RESEARCH:
   - Researcher at IKERLAN Research Center (MCC S. Coop.) 2001-2003
   - Member of Pattern Recognition & Speech Technology group in the University of the Basque Country since 2003
4. TEACHING:
   - University of the Basque Country since 2007
5. OTHERS
   - Researching stay in the Department of Information Systems and Computation. Technical University of Valencia. 03/2006?07/2006
   - Member of "International Speech Communication Association" (ISCA) and "Speech Technologies Thematic Network" (RTTH).
   - Member of the organizing committee of "AERFAI Summer School 2008"

6. AWARDS AND REVIEWS:

   - Special distinction given by Microsoft to the work "Different approaches to class-based language models using word segments" presented in IAPR "CORES 07"
   - Panelist in "AMBI-SYS 08"
   - Review of an article in "IEEE TASLP" journal.

7. PUBLICATIONS

   - R. Justo, M. I. Torres. Phrase classes in two-level language models for ASR. Pattern Analysis and Applications. (in press)
   - R. Justo, M. I. Torres. An approach to estimate perplexity values for language models based on phrase classes. Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis. Volume 5534 of LNCS, June 10-12 2009, Pvoa de Varzim (Portugal), pp 409–416
   - V. Guijarrubia, M. I. Torres, R. Justo. Morpheme-based Automatic Speech Recognition of Basque. Proceedings of the 4th Iberian Conference on Pattern Recognition and Image Analysis. Volume 5534 of LNCS, June 10-12 2009, Pvoa de Varzim (Portugal), pp 386–393
   - M. I. Torres, V. Guijarrubia, R. Justo, A. Pérez, F. Casacuberta. Statistical methods for speech technologies in basque language. Actas de las V Jornadas en Tecnologa del Habla. Bilbao 12-14 Noviembre, 2008.
   - R. Justo, O. Saz, V. Guijarrubia, A. Miguel, M. I. Torres, E. Lleida. Improving dialogue systems in a home automation environment. Proceedings of International Conference on Ambient Media and Systems (AMBI SYS 08). Quebec City, Canada. February 11-14, 2008.
   - R. Justo, M. I. Torres. Segment-based classes for language modeling within the field of CSR. Proceedings of the 12th Iberoamerican Congress on Pattern Recognition (CIARP). Valparaiso, Chile, November 13-16, 2007. Published in Volume 4756 of LNCS pp 714–723 .
   - A. Pérez, V. Guijarrubia, R. Justo, M. I. Torres, F. Casacuberta. A comparison of linguistically and statistically enhanced models for speech-to-speech machine translation. Proceedings of International Workshop on Spoken Language Translation (IWSLT 07). Trento, Italy. October 15-16, 2007
   - R. Justo, M. I. Torres: Different approaches to class-based language models using word segments. Proceedings of the IAPR International Conference on Computer Recognition Systems (CORES07). Wroclaw, Poland. October 22-25, 2007. Published in "Advances in Soft Computing". Volume 45, pp 421-428.
   - R. Justo, M. I. Torres. Two approaches to class-based language models for ASR. Proceeding of the IEEE Machine Learning for Signal Processing Workshop. Thessaloniki, Greece. August 27-29, 2007.
   - R. Justo, M. I. Torres. Phrases in category-based language models for Spanish and Basque ASR. Proceedings of Interspeech 2007. Antwerp, Belgium, August 27-31, 2007
   - R. Justo, M. I. Torres.Word segments in category-based language models for automatic speech recognition. Proceedings of the 3rd Iberian Conference on Pattern Recognition and Image Analysis. Volume 4477 of LNCS, June 6-8 2007, Girona (Spain), pp 249-256
   - R. Justo, M. I. Torres, Lluis Hurtado. Modelos de lenguaje basados en categoras semnticas en un sistema de dilogo de habla espontnea en castellano. Actas IV Jornadas en Tecnologa del Habla (IVJTH). Zaragoza. Noviembre 2006. ISBN: 84-96214-82-6
   - R. Justo, M. I. Torres, J.M. Benedí. Category-based language models in Spanish spoken dialogue systems. Procesamiento de Lenguaje Natural, vol. 37, pp 19-24 (SEPLN) Zaragoza, 13-15 Septiembre 2006.
   - J. M. Benedí, E. Lleida, A. Varona, M. J. Castro, I. Galiano, R. Justo, I. Lpez and A. Miguel. "Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA" Proceedings of LREC'06. Genova, Italy. 24-26 Mayo 2006.
   - R. Justo, M. I. Torres. Statistical and linguistic clustering for language modeling in ASR. Progress in Pattern Recognition, Image Anlisis and Applications. LNCS, vol 3773, pp 556-565. La Habana. Cuba. Noviembre 2005