Natural Language Processing techniques in Text-To-Speech synthesis and Automatic Speech Recognition

Alexandre Trilla

Departament de Tecnologies Mèdia
Enginyeria i Arquitectura La Salle (Universitat Ramon Llull), Barcelona, Spain
atrilla@salle.url.edu

January, 2009

Abstract

This working paper depicts the usage of Natural Language Processing techniques in the production of voice from an input text, *a.k.a.* Text-To-Speech synthesis, and the inverse process, which is the production of a written text transcription from an input voice utterance, *a.k.a.* Automatic Speech Recognition.

1 Introduction

Natural Language Processing (NLP) is the science most directly associated to processing human (natural) language. It derives from computer science since computers, or any other processing unit, are the target devices used to accomplish such processing. This description responds basically to the "Processing" particle in NLP. What makes NLP different from any other processing-related activity is the field of application: the human languages. They deal with more knowledge-related aspects thus requiring the support of learning capabilities by the processors of text.

In the end, it could be stated that most NLP or NLP-related computerized tasks can be wrapped by the more general concept of Machine Learning, clearly related to computer science, which contemplates any subject relating to the use of computers for the inference of the relevant features of the systems of study. Since the particular field of study is natural language, these learning techniques are of vital interest, because in some way, we humans make use of this kind of language as our basic means of communication and reasoning, inherently. If otherwise a formal language was to be studied (*e.g.*, a programming language), there would be no reason to make use of such learning approaches because the construction and logic issues bound to the formalism of that kind of language would already be known or predefined.

Common applications of sheer high-level NLP would deal solely with text data (at the input and output of the system) such as text classification, text summarization, question answering or machine translation. When approaching speech technologies other domains should be considered, despite NLP techniques refer exclusively to the textual analysis or synthesis of the applications. Either Text To Speech (TTS) synthesis or Automatic Speech Recognition (ASR) need a trustful module of NLP because text data always appears somewhere in the processing chain. TTS produces a speech utterance from an input text while ASR produces such text from an input speech utterance. Although these two objectives look very similar, their approach differs significantly.

This paper is centered on reviewing the main utilizations of NLP techniques for the two general speech technologies presented above. On one hand, [1] provides a generic text processing framework for English TTS synthesis, reviewing the tools needed to obtain a correct phonetic transcription of the input text. On the other hand, ASR applications are rather focused on the use of grammars (either hand-crafted or statistically derived) to construct Language Models as reviewed in [2], [3] and [4].

2 NLP for Speech Synthesis

TTS synthesis makes use of NLP techniques extensively since text data is first input into the system and thus it must be processed in the first place. [1] describes the different high-level modules involved in this sequential process:

Text Normalization Adapts the input text so as to be synthesized. It contemplates the aspects that are normally taken for granted when reading a text.
The **sentence segmentation** can be achieved though dealing with punctuation marks with a simple decision tree. But more confusing situations require more complex methods. Some examples of these difficulties are the period marking, the disambiguation between the capital letters in proper names and the beginning of sentences, the abbreviations, etc.
The **tokenization** separates the units that build up a piece of text. It normally splits the text of the sentences at white spaces and punctuation marks. This process is successfully accomlished with a parser.
Finally, **non-standard words** such as certain abbreviations (Mr., Dr., etc.), date constructs, phone numbers, acronyms or email and URL addresses need to be expanded into more tokens (units) in order to be synthesized correctly. Rules and diccionaries are of use to deal with non-standard words.

Part-of-Speech Tagging Assigns a word-class to each token. Thus this process consecutes the Text Normalization. Part-of-Speech taggers have to deal with unknown words (Out-Of-Vocabulary problem) and words with ambiguous POS tags (same structure in the sentence) such as nouns, verbs and adjectives. As an example, the use a participle as an adjective for a noun in “broken glass”.

Grapheme-to-Phoneme Conversion Assigns the correct phonetic set to the token stream. It must be stated that this is a continuous language dependent process since the phonetic transcriptions of the token boundaries are influenced by the transcriptions of the neighboring token boundaries. Thus, accounting for the influence of morphology and syllable structure can improve performace of Grapheme-to-Phoneme conversion [5].

Word Stress Assigns the stress to the words, a process tightly bound to the language of study. The phonological, morphological and word-class features are essential characteristics in this assignment: the stress is mostly determined by the syllable weight (phonological phenomena which treat certain syllable types as heavier than others [6]). See [1] for a wide set of references for this process.

The majority of the problems concerned by the above processes can be tackled either by rule-based or data-driven approaches, although this choice in some cases is clear due to the intractability of the resulting resolution for one of the two approaches. As an example, the traditional rule-based approach to perform phonetic transcriptions has proven to be very tedious and time consuming to develop and difficult to maintain, and in spite of its excellent transcription accuracy, this method has been put in the background in favor of computer learning approaches. See [7] for further details.

There are other NLP techniques related to speech synthesis that, although they are not essential, their use may improve the overall quality of the resulting speech, yielding a more natural impression to the users. Some of these approaches deal with emotions in order to produce affective, or expressive, speech synthesis. Aligned with this objective, the works presented in [8], [9] and [10] should be highlighted. These approaches are yet in an incipient stage and lots of research is being held presently as innovative solutions to attain such a natural interface.

3 NLP for Speech Recognition

Automatic Speech Recognition systems make use of NLP techniques in a fairly restricted way: they are based on grammars. This paper refers to a grammar as a set of rules that determine the structure of texts written in a given language by defining its morphology and syntax. ASR takes for granted that the incoming speech utterances must be produced according to this predetermined set of rules established by the grammar of a language, as it happens for a formal language. In that case, Context-Free Grammars (CFG) play an important role since they are well-capable of representing the syntax of that language while being efficient at the analysis (parsing) of the sentences [11]. For this reason/restriction, such language cannot be considered natural. ASR systems assume though that a large enough grammar rule set enable any (strictly formal) language to be taken for natural.

NLP techniques are of use in ASR when modeling the language or domain of interaction in question. Through the production of an accurate set of rules for the grammar, the structures for the language are defined. These rules can either be 1) hand-crafted or 2) derived from the statistical analyses performed on a labelled corpus of data. The former implies a great deal of hard-work since this process is not simple nor brief because it has to represent the whole set of grammatical rules for the application. The latter is generally the chosen one because of its programming flexibility at the expense of a tradeoff between the complexity of the process, the accuracy of the models and the volume of training and test data available (notice that the corpus has to be labelled, which implies a considerably hard workload). Since hand-crafted grammars depend solely on linguistics for a particular language and application they have little interest in machine learning research in general. Thus, the literature is extensive on the data-driven approaches (N-gram statistics, word lattices, etc.) bearing in mind that by definition a grammar-based representation of a language is a subset of a natural language.

Aiming at a flexible enough grammar to generalize the most typical sentences for an application, [2] and [3] end up building N-gram language models. N-grams model a language through the estimates of sequences of N consecutive words. While the former tackles the problem with a binary decision tree, the latter chooses to use more conventional Language Modeling theory (smoothing, cutoffs, context cues and vocabulary types). [11] also makes use of N-gram structures but it pursues a unified model integrating CFGs. Refer to the cited articles for further information. Lastly, [4] presents a means of dealing with spontaneous-speech through the spotlighting addition of automatic summarization including indexing, which extracts the gist of the speech transcriptions in order to deal with Information Retrieval (IR) and dialogue system issues.

Delving into the NLP-IR thread [12], ontologies have a deal of interest for their ability to construct Knowledge Bases in order to obtain some reasoning [13]. If merged with the present appeal for semantic-friendly interfaces [14], the resulting IR technology stands for one significant research topic.

4 Conclusions

This paper attempts to review the state-of-the-art Natural Language Processing techniques applied to speech technologies, specifically to Text-To-Speech synthesis and Automatic Speech Recognition. In

TTS (Section 2) the importance of NLP in processing the input text to be synthesized is reflected. The naturalness of the speech utterances produced by the signal-processing modules are tightly bound to the performance of the previous text-processing modules. In ASR (Section 3) the use of NLP particularly is complementary. It simplifies the recognition task by assuming that the input speech utterances must be produced according to a predefined set of grammatical rules. Its capabilities can though be enhanced through the usage of NLP aiming at more natural interfaces with a certain degree of knowledge. [15] reviews the major approaches proposed in language model adaptation in order to profit from this specific knowledge.

References

[1] U. D. Reichel and H. R. Pfitzinger, “Text preprocessing for speech synthesis,” in *Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation*, (Barcelona, Spain), pp. 207–212, June 2006.

[2] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, “A tree-based statistical language model for natural language speech recognition,” in *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, Issue 7, (Yorktown Heights, NY, USA), pp. 1001–1008, July 1989.

[3] P. Clarkson and R. Rosenfeld, “Statistical language modeling using the cmu-cambridge toolkit,” in *Proceedings EUROSPEECH* (N. F. G. Kokkinakis and E. Dermatas, eds.), vol. 1, (Rhodes, Greece), pp. 2707–2710, September 1997.

[4] S. Fururi, “Spontaneous speech recognition and summarization,” in *The Second Baltic Conference on HUMAN LANGUAGE TECHNOLOGIES*, (Tallinn, Estonia), pp. 39–50, April 2005.

[5] U. D. Reichel and F. Schiel, “Using morphology and phoneme history to improve grapheme-to-phoneme conversion,” in *Proceedings Eurospeech*, (Lisbon, Portugal), pp. 1937–1940, 2005.

[6] W. S. Allen, *Accent and Rhythm*. Cambridge University Press, 1973.

[7] F. Ivon, “Self-learning techniques for grapheme-to-phoneme conversion,” in *Onomastica Research Colloquium*, (London, UK), December 1994.

[8] C. Strapparava and R. Mihalcea, “Learning to identify emotions in text,” in *SAC’08: Proceedings of the 2008 ACM symposium on Applied computing*, (New York, NY, USA), pp. 1556–1560, ACM, 2008.

[9] V. Francisco and R. Hervás, “EmoTag: Automated Mark Up of Affective Information in Texts,” in *Proceedings of the Doctoral Consortium in EUROLAN 2007 Summer School* (C. Forascu, O. Postolache, G. Puscasu, and C. Vertan, eds.), (Iasi, Romania), pp. 5–12, July–August 2007.

[10] D. García and F. Alías, “Emotion identification from text using semantic disambiguation,” in *Procesamiento del Lenguaje Natural*, no. 40, pp. 75–82, March 2008.

[11] Y.-Y. Wang, M. Mahajan, and X. Huang, “A unified context-free grammar and n-gram model for spoken language processing,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. III, (Istanbul, Turkey), pp. 1639–1642, Institute of Electrical and Electronics Engineers, Inc., 2000.

[12] L. Zhou and D. Zhang, “NLPIR: a theoretical framework for applying natural language processing to information retrieval,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 54, no. 2, pp. 115–123, 2003.

[13] D. Bianchi and A. Poggi, “Ontology based automatic speech recognition and generation for human-agent interaction,” in *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2004. WET ICE 2004. 13th IEEE*

International Workshops on, (Modena, Italy), pp. 65–66, June 2004.

[14] J. Tejedor, R. Garca, M. Fernndez, F. J. Lpez-Colino, F. Perdrix, J. A. Macas, R. M. Gil, M. Oliva, D. Moya, J. Cols, , and P. Castells, "Ontology-based retrieval of human speech," in *Database and Expert Systems Applications, 2007. DEXA '07. 18th International Conference on*, (Regensburg, Germany), pp. 485–489, September 2007.

[15] J. R. Bellegarda, "Statistical language model adaptation: Review and perspectives," vol. 42, no. 1, pp. 93–108, 2004.