A Framework for Language-Independent Analysis and Prosodic Feature Annotation of Text Corpora

Dimitris Spiliotopoulos, Georgios Petasis and Georgios Kouroupetroglou

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Panepistimiopolis, Ilisia, GR-15784, Athens, Greece
dspiliot@di.uoa.gr, petasis@iit.demokritos.gr, koupe@di.uoa.gr

Abstract. Concept-to-Speech systems include Natural Language Generators that produce linguistically enriched text descriptions which can lead to significantly improved quality of speech synthesis. There are cases, however, where either the generator modules produce pieces of non-analyzed, non-annotated plain text, or such modules are not available at all. Moreover, the language analysis is restricted by the usually limited domain coverage of the generator due to its embedded grammar. This work reports on a language-independent framework basis, linguistic resources and language analysis procedures (word/sentence identification, part-of-speech, prosodic feature annotation) for text annotation/processing for plain or enriched text corpora. It aims to produce an automated XML-annotated enriched prosodic markup for English and Greek texts, for improved synthetic speech. The markup includes information for both training the synthesizer and for actual input for synthesising. Depending on the domain and target, different methods may be used for automatic classification of entities (words, phrases, sentences) to one or more preset categories such as "emphatic event", "new/old information", "second argument to verb", "proper noun phrase", etc. The prosodic features are classified according to the analysis of the speech-specific characteristics for their role in prosody modelling and passed through to the synthesizer via an extended SOLE-ML description. Evaluation results show that using selectable hybrid methods for part-of-speech tagging high accuracy is achieved. Annotation of a large generated text corpus containing 50% enriched text and 50% canned plain text produces a fully annotated uniform SOLE-ML output containing all prosodic features found in the initial enriched source. Furthermore, additional automatically-derived prosodic feature annotation and speech synthesis related values are assigned, such as word-placement in sentences and phrases, previous and next word entity relations, emphatic phrases containing proper nouns, and more.

1 Introduction

Text annotation is a procedure where certain meta-information gets identified and associated with the entities in a text corpus. Such information is commonly used in computational linguistics for language analysis, speech processing, natural language processing, speech synthesis, and other areas. The type of information that is analyzed and associated to text units may span the linguistic analysis tree (grammatical, syntactic,

Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): TSD 2008, LNAI 5246, pp. 517–524, 2008.

morphological, semantic, pragmatic, phonological, phonetic), as well as include any other description that may be of use.

Speech synthesizers traditionally perform a part-of-speech analysis and build the syntactic tree of the text in order to assign prosody [1]. General purpose Text-to-Speech (TtS) systems use certain language processing subsystems, such as sentence segmentation and part-of-speech tagging, for the analysis of the written text input. Depending on the actual system, such analysis may suffer from inherent statistical error accuracy that may be due to the design and implementation of the respective modules or language ambiguity. However, TtS systems may employ language analysis modules that are designed for high accuracy in specific thematic domains for which they seem to perform adequately. The respective accuracy when used for generic or other thematic domains may fall under unacceptable levels. Additionally, the language processing modules embedded in TtS systems are not usually designed to identify and extract higher-level linguistic information, such as semantic or pragmatic factors, that may be used to aid speech synthesis.

Concept-to-Speech (CtS) systems seem to provide an ideal means of text analysis. The Natural Language Generator (NLG) component of the CtS systems produces processed and annotated text as input for the speech synthesis module [2]. The NLG output text is generated as error-free syntactically annotated text exhibiting full disambiguation. In addition, further linguistic information may be generated providing considerable aid to guide synthesis. CtS systems, as a result, utilize the linguistic features from the natural language generation phase in order to produce significantly improved synthesized speech [3]. One of the major drawbacks of CtS systems is that the NLGs are designed to operate in specific thematic domains, and thus restricted to limited domain text generation. To make the things more complicated, the text output may not always be generated by the system. There may exist chunks of plain unprocessed text (canned) designed to be included in the output. These include groups of words, phrases or whole sentences that contain language that is too complicated for the NLG to fully generate. Such example is the MPIRO corpus [4] where more than 40% of the text descriptions of a museum exhibit domain is canned text. A linguistic analysis of that portion of text can provide a fully analyzed, uniformly annotated corpus, an essential and important benefit for speech synthesis.

Previous works that have explored natural language generated texts show that linguistically enriched annotated text input to a speech synthesizer can lead to improved naturalness of speech output [5,6]. Generation of tones and prosodic phrasing from high level linguistic input produces better prosody than plain texts do [7]. When such input can be provided, the language processing from the TtS system can be superseded.

In this work, a language-independent framework for language analysis and semantic annotation is presented. The aim is to produce uniform enriched text description, similar to the one generated by the natural language generation component of a CtS system, starting from plain or partially annotated text whether that may come from a natural language generator or a plain text document. This framework has been used successfully for the design, implementation and evaluation of a methodology for automatic annotation of large domain-dependent Greek text corpora.

This work reports on the set of linguistic features and information that needs to be considered and the description of the workflow and key modules that are employed for enriched text annotation for English and Greek text. Furthermore, the nature of the text analysis and prosodic feature incorporation are explored for focus prominence calculation for synthetic speech.

2 Enriched Text Annotation

TtS systems generally accept plain (or "raw") text as input, using specialized algorithms to internally generate the needed natural language data prior to synthesis. However, the algorithms that are usually implemented for such tasks are not powerful enough to broadly identify additional information about several linguistic phenomena from the plain text form, thus limiting the depth of text analysis and the derived description. A valuable alternative is to use pre-processed annotated text as input to the speech synthesizer. Enriched text of that kind exhibits major advantage over plain text as it retains structural and discourse level information. Each of the above types of linguistic information is described by sets of features that can be used to generate improved prosody in speech synthesis. Depending on the domain as well as the type of text, different sets of features may be used for maximum improvement.

As an alternative to generated text, existing plain text can be adequately processed to derive annotated NLG-similar output, essentially gaining advantage for the prosody modelling stage in speech synthesis. In order to do that efficiently, automated analysis and annotation should be made available for the most language analysis stages. A breakdown of the identifiable distinct processes is:

- Word/Sentence identification and segmentation.
- Morphological analysis (part-of-speech tagging and noun-phrase identification).
- Calculation/annotation of prosodic features.
- Creation/export to appropriate XML format description for speech synthesis.

As described in the following paragraphs, fully automated analysis can be achieved for all processes. The enriched linguistic annotation needs to be exported to a well-tested and reliable standard markup, such as XML. All the above processes have been implemented through the utilisation of the Ellogon Language Engineering Platform [8] platform and implemented the speech-oriented natural language analysis and annotation components [9].

As shown in Figure 1, the input may be either fully or partially annotated text (e.g. from a Natural Language Generator) or plain unformatted text. Information from the enriched input is extracted and used for the annotation of the plain text. The prosodic feature annotation assigns prosodically important values for the calculation of intonation focus for higher quality speech synthesis.

3 Morphological Analysis

Pre-processing mainly includes word and sentence identification, as well as part-of-speech (POS) tagging. For English texts, a POS tagger based on machine learning is

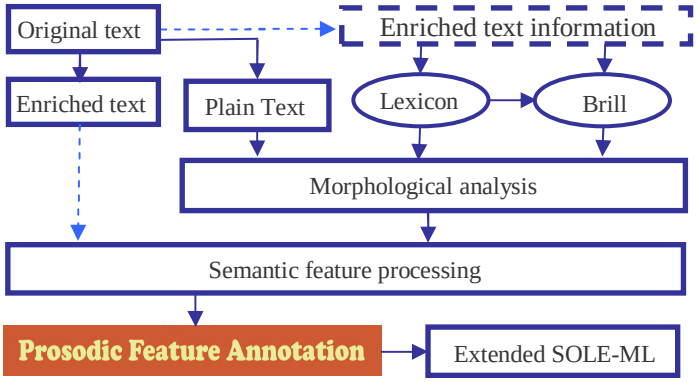

Fig. 1. The annotation workflow

used, while for Greek texts a combination of lexicon-based and machine learning analysis is preferred. Word and sentence identification are performed by a rule-based component (HTokenizer) that presents an accuracy that approaches 100% for both languages. For part-of-speech tagging, the implementation is based on Transformation-based Error-driven learning [10] and provides models for English with an accuracy that approaches 97% measured as average of several accuracy measurements performed on various thematic domains.

For Greek, the common approach for most embedded systems is the use of Lexicon-based POS taggers. This approach is used by most speech synthesisers and yields accuracy between 75–85% depending on the domain of the text corpora. This low accuracy in most cases hinders poor final prosody prediction. This is due to Greek being an inflectional language with vast vocabulary that cannot be covered by lexicons. In order to increase the accuracy of POS tagging when processing documents in the Greek language, we used a hybrid approach, a combination of a lexicon-based POS tagger and a rule-based (Brill) POS tagging component. Two morphological lexicons for the Greek language have been combined in order to build a lexicon-based POS tagger with the highest possible coverage. The first lexicon is a large-scale morphological lexicon for the Greek language, developed exclusively for the system [11]. The lexicon consists of ∼60,000 lemmas that correspond to ∼710,000 different word forms (Greek is an inflexional language). The second lexicon is property of the Speech Group, University of Athens, used in the DEMOSTHeNES speech composer [12] and contains ∼60,000 lemmas, which correspond to ∼650,000 word forms. Both lexicons yield a word form identification span of ∼880,000.

The hybrid approach was applied to the full generated corpus using two different ways in order to examine and evaluate the best procedure:

In the first approach, the built-in POS tagger and the lexicon-based POS tagger are both applied independently. Depending on the actual corpus and relative precision of the lexicon and HBrill modules, a word can be set to be assigned a value by either tagger (or both). The default state is that if a word contained in any of the two lexicons and thus is assigned a POS category by the lexicon-based tagger, this categorization becomes the

final POS of the word, ignoring any categorization performed by the machine learning POS tagger. On the other hand, if a word is not found in any of the two lexicons, the categorization presented by the built-in POS tagger is assigned. The machine learning based POS tagger uses an extension of the Penn Tree Bank tagset, which contains additional information regarding number and gender of words [13]. This approach achieved an accuracy of 95%.

The second approach sees that the Lexicon-based component is always followed be the machine learning POS tagger. Initial values are extrapolated from the lexicons and fed as initial states for the machine learning algorithm which provides the final value. In the case of partially annotated texts, the values of the pre-annotated word tokens were used for initial values in similar word forms since they were 100% correct coming from the natural language generator. This approach yielded total accuracy $>97\%$ for plain and $>98\%$ for partially annotated Greek texts and was the preferred choice.

4 Prosodic Feature Annotation

Previous research shows that higher-level linguistic information such as semantic features can be used to improve prosody modelling for speech synthesis [6]. This is because part-of-speech and phrase type information alone cannot always infer certain intonational focus points since those are not only affected by syntax but also by semantic and pragmatic factors [14]. For prosody modelling in speech synthesis, these factors can be used for calculation, deduction and verification of focus prominence and are accounted for by enriching the text corpus accordingly.

In our corpus, the plain text was annotated using the hybrid part-of-speech technique. Then, the results were validated and updated using the part-of-speech information from the enriched corpus. The benefit is twofold, the values are checked with the correct ones from the enriched text (if such is available for a lexical item) and key items are assigned specific values where appropriate. After that, certain semantic factors are calculated and added to the meta-information pool

Figure 2 shows an example of how semantic factors such as *newness (new or old infromation)*, *contrast*, *explicit emphasis*, *first or second argument to verb* may be used for determining intonational focus prominence.

The intonational focus is assigned in a scale of three, strong focus ‘3’, normal focus ‘2’, and weak focus ‘1’. The features in bold are the ones computed from the information provided by the enriched portion of the text. Although *newness* is a key factor for strong intonational focus, certain validation checks in the algorithm make sure that only the proper lexical items are assigned. Validation factors are proper-noun and second-argument-to-verb (arg2) as well as explicit factors such as *emphasis* and *contrast*. As a result, strong focus ‘3’ is assigned when validation factors arg2 and/or proper-noun exist for a new information (e.g., #1-2) while old information (e.g. #8-9) gets weak focus, as shown below:

Strong focus prominence: newness_TRUE (validation=passed)
Normal focus prominence: newness_FALSE (validation=passed)
Weak focus prominence: newness_TRUE (validation=failed)
No focus prominence: newness_FALSE (validation=failed)

However, it can be seen that explicit factors elevate the focus prominence, clearly providing explicit emphatic events as in the case of *splachnoscopy* (# 8) where if it were not for the *emphasis* factor it would have been assigned weak focus since it is an already given piece of information.

This exhibit is an amphora, created during the archaic period. It dates from the early fifth century before Christ. It was found in Beotea but it was made in Athens. It depicts a warrior performing splachnoscopy before leaving for battle. Splachnoscopy is the study of animal entrails, through which people tried to predict the future. It was one of the most common divination methods used in the archaic period. This amphora was painted by the painter of Kleofrades and was decorated with the red figure technique.

#	Lexical item	Focus	Prosodic features
1	amphora	3	[newness_TRUE, arg2]
2	archaic period	3	[newness_TRUE, arg2]
3	Christ	2	[newness_FALSE, proper-noun]
4	It was … in Athens	-	[contrast]
5	Beotea	3	[newness_TRUE, arg2, ***proper-noun***]
6	Athens	3	[newness_TRUE, arg2, ***proper-noun***]
7	splachnoscopy	3	[newness_TRUE, arg2, ***emphasis***]
8	Splachnoscopy	3	[newness_FALSE, arg1, ***emphasis***]
9	archaic period	1	[newness_FALSE, arg2]
10	amphora	1	[newness_FALSE, arg2]
11	the painter of Kleofrades	3	[newness_TRUE, arg2, ***proper-noun***]
12	red figure	3	[newness_TRUE, arg2, ***proper-noun***]

Fig. 2. Focus prominence identification from semantic factors

Contrast is a rather generalized annotation that was implemented as a rule in the process and was initiated due to the fact that domain contained several instances of similarly NLG-derived phrases. The rule applies to both Greek and English text and elevates the main verb(s) and the conjunction to a mid-level emphasis, thus assigning explicitly a normal focus prominence marker (not showing in Figure 2).

From the above, it is obvious that the precision of part-of-speech identification is quite important since certain lexical items are validated for their assigned focus prominence using the part-of-speech information against the identified prosodic features.

5 SOLE-ML Description

The enriched text meta-information is encoded using an open XML schema. It is an extension (to cater for the semantic/prosodic description) of the SOLE-ML description [15], and was originally built as an annotation scheme for CtS synthesis, used as markup for the enriched text output of the ILEX generator [2]. It has been successfully used in earlier works, a well-tested means of representing enriched linguistic information, and is now standard input of the DEMOSTHeNES speech composer. The automatic extraction to the extended XML description based on SOLE-ML encodes all prosodic features. Figure 3 shows the XML output for the sentence *"It was found in Beotea but it was made in Athens."* from the text paragraph shown in Figure 2.

```
<utterance>
<relation name="Word" structure-type="list">
<wordlist>
<w id="w20">It</w>
<w id="w21">was</w>
<w id="w22">found</w>
<w id="w23">in</w>
<w id="w24">Beotea</w>
<w id="w25">but</w>
<w id="w26">it</w>
<w id="w27">was</w>
<w id="w28">made</w>
<w id="w29">in</w>
<w id="w30" punct=".">Athens</w>
</wordlist>
</relation>
<relation name="Group" structure-type="list">
</relation>
<relation name="Syntax" structure-type="tree">
<elem phrase-type="S">
<elem phrase-type="prosody" contrast>
<elem lex-cat="PRONOUN" href="w#id(w20)"/elem>
<elem phrase-type="prosody" mid-emphasis-verb>
<elem lex-cat="VERB" href="w#id(w21)"/elem>
<elem lex-cat="VERB" href="w#id(w22)"/elem>
</elem>
<elem lex-cat="PREPOS" href="w#id(w23)"/elem>
<elem phrase-type="prosody" newness="true", arg2, proper-noun>
<elem lex-cat="NOUN" href="w#id(w24)"/elem>
</elem>
<elem phrase-type="prosody" mid-emphasis-conj>
<elem lex-cat="CONJNCT" href="w#id(w25)"/elem>
</elem>
<elem lex-cat="PRONOUN" href="w#id(w26)"/elem>
<elem phrase-type="prosody" mid-emphasis-verb>
<elem lex-cat="VERB" href="w#id(w27)"/elem>
<elem lex-cat="VERB" href="wl#id(w28)"/elem>
</elem>
<elem lex-cat="PREPOS" href="w#id(w29)"/elem>
<elem phrase-type="prosody" newness="true", arg2, proper-noun >
<elem lex-cat="NOUN" href="w#id(w30)"/elem>
</elem>
</elem>
</elem>
</relation>
</utterance>
```

Fig. 3. The XML description

A wordlist of all tokens (words) and punctuation values takes up the first part (<*wordlist*>), followed by the syntax tree, prosodic features, and other high-level information (<relation>). This is the input for the speech synthesizer.

6 Evaluation and Discussion

The proposed framework utilises the meta-information contained in enriched automatically generated texts in order to compute and annotate both the enriched and the plain text with prosodic features that aid focus prominence in synthetic speech. The uniformly annotated target text contains enough elements to aid focus prominence using the modified speech synthesizer for Greek or an equivalent for English. An evaluation of the performance of the hybrid morphological analysis methods was performed for both Greek and English texts, shown in Table 1.

Table 1. Plain text part-of-speech annotation

Corpus (plain text)		Lexicon	Brill	Hybrid
English	precision	0.90	0.97	0.98
	recall	0.77	0.92	0.98
Greek	precision	0.88	0.94	0.98
	recall	0.75	0.84	0.92

These results include the prime importance validation factor in our approach *proper-noun,* while exclude all other features that are calculated later in the process.

Enriched text annotation using naturally generated meta-information for a specific domain greatly enhances the intonational focus prominence predictors of a speech synthesizer. A strong indication of focus based on the new or already given information validated by the type of the lexical item works exceptionally well for domain-dependent corpora where the prosodic features can be more easily calculated automatically. This

leads to enhanced input for speech synthesis, while bypassing all internal language analysis modules of the synthesizer, results on improved prosody prediction.

Acknowledgements. The work described in this paper has been funded by the European Social Fund and Greek National Resources under the RHETOR project of the Information Society programme, Hellenic General Secretariat of Research and Technology.

References

1. Taylor, P., Black, A., and Caley, R., "The architecture of the festival speech synthesis system," Proc. 3rd ESCA Workshop on Speech Synthesis, Australia, pp.147–151, 1998.
2. O'Donnel, M., Mellish, C., Oberlander, J., and Knott, A., "ILEX: An architecture for a dynamic hypertext generation system", Natural Language Engineering, vol.7, no.3, pp. 225–250, 2001.
3. Hitzeman, J., Black, A., Taylor, P., Mellish, C., and Oberlander, J. "On the Use of Automatically Generated Discourse-Level Information in a Concept-to-Speech Synthesis System", Proc. 5th Int. Conf. on Spoken Language Generation (ICSLP): 2763–2768, 1998.
4. Isard, A., Oberlander, J., Androutsopoulos, I., and Matheson, C., "Speaking the Users' Languages". *IEEE Intelligent Systems*, 18(1):40–45, 2003.
5. Pan, S., McKeown, K., and Hirschberg, J., "Exploring features from natural language generation for prosody modeling" Computer Speech and Language, 16:457–490, 2002.
6. Xydas, G., Spiliotopoulos, D., and Kouroupetroglou, G., "Modeling Improved Prosody Generation from High-Level Linguistically Annotated Corpora". IEICE Trans. of Inf. and Syst., Special Section on "Corpus-Based Speech Technologies", vol. E88-D, no 3, March 2005, pp. 510–518.
7. Black, A., and Taylor, P., "Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input" Proc. 3rd Int. Conf. on Spoken Language Processing, pp.715–718, Yokohama, Japan, 1994.
8. Petasis, G., Karkaletsis, V., Paliouras, G., Androutsopoulos, I., and Spyropoulos, C.D., "Ellogon: A New Text Engineering Platform". Proc. 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002), pp. 72–78, Las Palmas, Canary Islands, Spain, May 2002.
9. Ellogon Language Engineering Platform, Speech tools add-ons, `http://www.ellogon.org/speech/`.
10. Brill, E., "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging". Computational Linguistics, 21:543–565, 1995.
11. Petasis, G., Karkaletsis, V., Farmakiotou, D., Androutsopoulos, I., and Spyropoulos, C.D., "A Greek Morphological Lexicon and its Exploitation by Natural Language Processing Applications". Lecture Notes on Computer Science, vol.2563, Springer Verlag, 2003.
12. Xydas, G. and Kouroupetroglou, G., "The DEMOSTHeNES Speech Composer", Proc. 4th ISCA Workshop on Speech Synthesis, Perthshire, Scotland, pp.167–172, 2001.
13. Petasis, G., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D., and Androutsopoulos, I., "Resolving Part-of-Speech Ambiguity in the Greek Language Using Learning Techniques". Fakotakis et al. (Eds.), Machine Learning in Human Language Technology, pp. 29–34, 1999.
14. Bolinger, D., Intonation and its Uses: Melody in grammar and discourse, Edward Arnold, London, 1989.
15. Hitzeman, J., Black, A., Mellish, C., Oberlander, J., Poesio, M., and Taylor, P., "An annotation scheme for Concept-to-Speech synthesis", Proc. 7th European Workshop on Natural Language Generation, Toulouse France, pp. 59–66, 1999.