SPARSAR: An Expressive Poetry Reader

[name]
[address]
[address]
[e-mail]

Abstract

We present SPARSAR, a system for the automatic analysis of poetry(and text) style which makes use of NLP tools like tokenizers, sentence splitters, NER (Name Entity Recognition) tools, and taggers. In addition the system adds syntactic and semantic structural analysis and prosodic modeling. We do a dependency mapping to analyse the verbal complex and determine Discourse Structure. Another important component of the system is a phonological parser to account for OOVWs, in the process of grapheme to phoneme conversion of the poem. We also measure the prosody of the poem by associating mean durational values in msecs to each syllable from a database of syllable durations; to account for missing syllables we built a syllable parser with the aim to evaluate durational values for any possible syllable structure. A fundamental component for the production of emotions is the one that performs affective and sentiment analysis. This is done on a line by line basis. Lines associated to specific emotions are then marked to be pronounced with special care for the final module of the system, which is reponsible for the production of expressive reading by a TTS module, in our case the one made available by Apple on their computers. Expressive reading is allowed by the possibility to interact with the TTS.

1 Introduction

We present SPARSAR, a system for poetry (and text) style analysis by means of parameters derived from deep poem (and text) analysis. We use our system for deep text understanding called VENSES(XXX,2005) for that aim. SPARSAR(XXX,2013a) works on top of the output provided by VENSES and is organized in three main modules which can be used also to analyse similarities between couples of poems by the same or different poet and similarities between collections of poems by a couple of poets. In addition to what is usually needed to compute text level semantic and pragmatic features, poetry introduces a number of additional layers of meaning by means of metrical and rhyming devices. For these reasons more computation is required in order to assess and evaluate the level of complexity that a poem objectively contains. We use prosodic durational parameters from a database of English syllables we produced for a prosodic speech recognizer (XXX,1990). These parameters are used to evaluate objective presumed syllable and feet prosodic distribution at line level. The sum of all of these data is then used to create a parameterized version of the poem to be read by a TTS, with an appropriate expressivity. Expressive reading is generated by combining syntactic, semantic, lexical and prosodic information. It is a well-known fact that TTS systems are unable to produce utterances with appropriate prosody(van Santen et al.,2003)[1]. Besides the general problems related to TTS reading normal texts, when a poem is inputted to the TTS the result is worsened by the internal rules which compute stanza boundaries as sentence delimiters. So every time there are continuations or enjambements from one stanza to the next the TTS will not be able to see it, and will produce a long pause. The TTS is also blind to line boundaries. More importantly, the TTS reads every sentence with the same tone, thus contributing an unpleasant repeated overall boring sense which does not correspond to the contents read. This is why sentiment analysis can be of help, together with semantic processing at discourse level.

As regards affective or emotional reading, then, the prosody of current TTS systems is neutral, and generally uses flat intonation contours. Producing "expressive" prosody will require modifying rhythm, stress patterns and intonation as described in section 4(see Kao & Jurafsky,2012).

[1] as he puts it, "The wrong words are emphasized, phrase boundaries are not appropriately indicated, and there is no prosodic structure for longer stretches of speech. As a result, comprehension is difficult and the overall listening experience is disconcerting…" (ibid.,1657).

The paper is organized as follows: here below a subsection contains a short state of the art limited though to latest publications; section 2 shortly presents SPARSAR; section 3 is dedicated to Prosody, Rhyming and Metrical Structure; a short state of the art of expressive reading is presented in section 4, which is devoted to TextTo-Speech and parameters induction from the analysis. Eventually we present an evaluation, a conclusion and work for the future.

2 PARSAR - Automatic Analysis of Poetic Structure and Rhythm with Syntax, Semantics and Phonology

SPARSAR[8] produces a deep analysis of each poem at different levels: it works at sentence level at first, than at line level and finally at stanza level. The structure of the system is organized as follows: at first syntactic, semantic and grammatical functions are evaluated. Then the poem is translated into a phonetic form preserving its visual structure and its subdivision into lines and stanzas. Phonetically translated words are associated to mean duration values taking into account position in the word and stress. Taking into account syntactic and semantic information, we then proceed to "demote" word stress of dependent or functional words. At the end of the analysis of the poem, the system can measure the following parameters: mean verse length in terms of msec. and in number of feet. The latter is derived by a line and stanza representation of metrical structure. More on this topic below.

Another important component of the analysis of rhythm is constituted by the algorithm that measures and evaluates rhyme schemes at stanza level and then the overall rhyming structure at poem level. As regards syntax, we build chunks and dependency structures. To complete our work, we introduce semantics at two levels. On the one hand, we isolate verbal complex in order to verify propositional properties, like presence of negation, computing factuality from a crosscheck with modality, aspectuality – that we derive from our lexica – and tense. We also classify referring espressions by distinguishing concrete from abstract nouns, identifying highly ambiguous from singleton concepts (from number of possible meanings from WordNet and other similar repositories). Eventually, we carry out a sentiment analysis of every poem, thus contributing a three-way classification: neutral, negative, positive that can be used as a powerful tool for expressive purposes.

3 Rhetoric Devices, Metrical and Prosodic Structure

The second module takes care of rhetorical devices, metrical structure and prosodic structure. This time the file is read on a line by line level by simply collecting strings in a sequence and splitting lines at each newline character. In a subsequent loop, whenever two newlines characters are met, a stanza is computed. In order to compute rhetorical and prosodic structure we need to transform each word into its phonetic counterpart, by accessing the transcriptions available in the CMU dictionary. The Carnegie Mellon Pronouncing Dictionary is freely available online and includes American English pronunciation[2]. We had available a syllable parser which was used to build the VESD database of English syllables (XXX, 1999a) (Venice English Syllable Database) to be used in the Prosodic Module of SLIM, a system for prosodic self-learning activities(XXX,2010), which we use whenever we have a failure of our pronunciation dictionary which covers some 170,000 entries.

Remaining problems to be solved are related to ambiguous homographs like "import" (verb) and "import" (noun) and are treated on the basis of their lexical category derived from previous tagging; and Out Of Vocabulary Words (OOVW). If a word is not found in the dictionary, we try different capitalizations, as well as breaking apart hyphenated words, and then we check with simple heuristics, differences in spelling determined by British vs. American pronunciation. Then we proceed by morphological decomposition, splitting at first the word from its prefix and if that still does not work, its derivational suffix. As a last resource, we use an orthographically based version of the same dictionary to try and match the longest possible string in coincidence with our OOVW. Some words we had to reconstruct are: wayfare, gangrened, krog, copperplate, splendor, filmy, seraphic, unstarred, shrive, slipstream, fossicking, unplotted, corpuscle, thither, wraiths, etc. In some cases, the problem that made the system fail was the syllable which was not available in our database of syllable durations, VESD[3]. This problem has been coped with

[2] It is available online at <http://www.speech.cs.cmu.edu/cgi-bin/cmudict/>.

[3] In VESD, syllables have been collected from WSJCAM, the Cambridge version of the continuous speech recognition corpus produced from the Wall Street Journal, distributed by the Linguistic Data Consortium (LDC). We worked on a subset of 4165 sentences, with 70,694 words which consti-

by launching the syllable parser and then computing durations from the component phonemes, or from the closest similar syllable available in the database. We only had to add 12 new syllables for a set of approximately 500 poems that we computed to test the system.

3.1 Computing Metrical Structure and Rhyming Scheme

Any poem can be characterized by its rhythm which is also revealing of the poet's peculiar style. In turn, the poem's rhythm is based mainly on two elements: meter, that is distribution of stressed and unstressed syllables in the verse, presence of rhyming and other poetic devices like alliteration, assonance, consonance, enjambements, etc. which contribute to poetic form at stanza level.

We follow Hayward (1991) to mark a poetic foot by a numerical sequence that is an alternation of 0/1: "0" for unstressed and "1" for stressed syllables. The sequence of these sings makes up the foot and depending on number of feet one can speak of iambic, trochaic, anapestic, dactylic, etc. poetic style. But then we deepen our analysis by considering stanzas as structural units in which rhyming plays an essential role. Secondly we implement a prosodic acoustic measure to get a precise definition of rhythm. Syllables are not just any combination of sounds, and their internal structure is fundamental to the nature of the poetic rhythm that will ensue. The use of duration has allowed our system to produce a model of a poetry reader that we implement by speech synthesis. To this aim we assume that syllable acoustic identity changes as a function of three parameters:

- internal structure in terms of onset and rhyme which is characterized by number consonants, consonant clusters, vowel or diphthong
- position in the word, whether beginning, end or middle
- primary stress, secondary stress or unstressed

4 TTS and Modeling Poetry Reading

The other important part of the work regards using the previous analyses to produce intelligible, correct, appropriate and possibly pleasant or catchy poetry reading by a TextToSpeech system. In fact, the intention was more ambitious and was producing an "expressive" reading of a poem in the sense also intended by work reported in Ovesdotter & Sprout(2005), Ovesdotter(2005), Scherer(2003). In Ovesdotter & Sprout(2005), the authors present work on fairy tales, intended to use positive vs negative classification of sentences to produce a better reading. To that aim they used a machine learning approach, based on the manual annotation of some 185 children stories[4]. They reported accuracy results around 63% and F-score around 70%, which they explain may be due to a very low interannotator agreement, and to the fact that the dataset was too small. In Ovesdotter(2005) the author presents work on the perception of emotion based again on fairy tales reading by human readers. The experiment had the goal of checking the validity of the association of acoustic parameters to emotion types. Global acoustic features included F0, intensity, speech rate in number of words, feet, syllables per minute, fluency, i.e. number of pauses or silences. The results show some contradictory data for ANGRY state, but fully compliant data for HAPPY[5]. These data must be regarded as tendencies and are confirmed by experiments reported also in Scherer(2003) and Schröder(2001). However, it must be underlined that all researchers confirm the importance of semantic content, that is the meaning as a means for transmitting affective states.

The TTS we are now referring to is the one freely available under Mac OSX in Apple's devices. In fact, the output of our system can be used to record .wav or .mpeg files that can then be played by any sound player program. The information made available by the system is sufficiently deep to allow for Mac TTS interactive program to adapt the text to be read and model it

tute half of the total number of words in the corpus amounting to 133,080. We ended up with 113,282 syllables and 287,734 phones. The final typology is made up of 44 phones, 4393 syllable types and 11,712 word types. From word-level and phoneme-level transcriptions we produced syllables automatically by means of a syllable parser. The result was then checked manually.

[4] Features used to learn to distinguish "emotional" from "neutral" sentences, include (ibid., 582): first sentence in the story; direct speech; thematic story type (animal tale, ordinary folk-tale, jokes and anecdotes); interrogative and exclamative punctuation marks; sentence length in words; ranges of story progress; percent of semantic words (JJ, N, V, RB); V count in sentence, excluding participles; positive and negative words; WordNet emotion words; interjections and affective words; content BOW: N,V,JJ,RB words by POS.

[5] In particular, "angry" was associated with "decreased F0" and "decreased speech rate", but also an increased "pausing". On the contrary, "happy" showed an "increased F0, intensity, pausing" but a "decreased speech rate". "Happy" is similar to "surprised", while "angry" is similar to "sad".

accurately. We used the internal commands which can modify sensibly the content of the text to be read. The voices now available are pleasant and highly intelligible. We produced a set of rules that take into account a number of essential variables and parameter to be introduced in the file to be read. Parameters that can be modified include: Duration as Speaking Rate; Intonation from first word marked to a Reset mark; Silence introduced as Durational value; Emphasis at word level increasing Pitch; Volume from first word marked to a Reset mark, increasing intensity. We discovered that Apple's TTS makes mistakes when reading some specific words, which we then had to input to the system in a phonetic format, using the TUNE modality.

The rules address the following information:

- the title
- the first and last line of the poem
- a word is one of the phonetically spelled out words
- a word is the last word of a sentence and is followed by an exclamation/interrogative mark
- a word is a syntactic head (either at constituency or dependency level)
- a word is a quantifier, or marks the beginning of a quantified expression
- a word is a SUBJect head
- a word marks the end of a line and is (not) followed by punctuation
- a word is the first word of a line and coincides with a new stanza and is preceded by punctuation
- a line is part of a sentence which is a frozen or a formulaic expression with specific pragmatic content specifically encoded
- a line is part of a sentence that introduces new Topic, a Change, Foreground Relevance as computed by semantics and discourse relations
- a line is part of a sentence and is dependent in Discourse Structure and its Move is Down or Same Level
- a discourse marker indicates the beginning of a subordinate clause

5 Evaluation, Conclusion and Future Work

We have done a manual evaluation by analysing a randomly chosen sample of 50 poems out of the 500 analysed by the system. The evaluation has been made by a secondary school teacher of English literature, expert in poetry[6]. We asked the teacher to verify the following four levels of analysis: 1. phonetic translation; 2. syllable division; 3. feet grouping; 4. metrical rhyming structure. Results show a percentage of error which is around 5% as a whole, in the four different levels of analysis. A first prototype has been presented in(XXX,2013a), and improvements have been done since then; but more work is needed to tune prosodic parameters for expressivity rendering both at intonational and rhythmic level. The most complex element to control seems to be variations at discourse structure which are responsible for continuation intonational patterns vs. beginning of a new contour.

Reference

XXX. 1999. "Prosodic Modeling for Syllable Structures from the VESD - Venice English Syllable Database", in *Atti 9° Convegno GFS-AIA*, Venezia, 161-168.

XXX. 2008. "Speech Synthesis for Language Tutoring Systems", in V.Melissa Holland & F.Pete Fisher(eds.), (2008), *The Path of Speech Technologies in Computer Assisted Language Learning*, Routledge - Taylor and Francis Group-, New York, 123-150.

XXX, 2010. "Prosodic tools for language learning", *International Journal of Speech Technology*. 12(4):161-184.

XXX, 2013a. SPARSAR: a System for Poetry Automatic Rhythm and Style AnalyzeR, SLATE 2013, Demonstration Track.

XXX. 2005. "VENSES – a Linguistically-Based System for Semantic Evaluation", in J. Quiñonero-Candela et al.(eds.), 2005. *Machine Learning Challenges*. LNCS, Springer, Berlin, 344-371.

M. Hayward. 1991. "A connectionist model of poetic meter". *Poetics*, 20, 303-317.

Justine Kao and Dan Jurafsky. 2012. "A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry". in *Proc. NAACL Workshop on Computational Linguistics for Literature*.

Cecilia Ovesdotter Alm, Richard Sproat, 2005. "Emotional sequencing and development in fairy tales", In *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction, ACII '05*.

Cecilia Ovesdotter Alm, 2005. "Emotions from text: Machine learning for text-based emotion prediction", In *Proceedings of HLT/EMNLP*, 347-354.

Jan van Santen, Lois Black, Gilead Cohen, Alexander Kain, Esther Klabbers,Taniya Mishra, Jacques de Villiers, and Xiaochuan Niu. 2003. "Applications of Computer Generated Expressive Speech for Communication Disorders", in *Proc. Eurospeech*, Geneva, 1657-1660.

K. R. Scherer. 2003. "Vocal communication of emotions: a review of research paradigms", *Speech Communication*, 40(1-2):227-256.

[6] I here acknowledge the contribution of XXX and thank her for the effort.