Text classification of domain-styled text and sentiment-styled text for expressive speech synthesis

Alexandre Trilla, Francesc Alías, Isaac Lozano

GTM – Grup de Recerca en Tecnologies Mèdia
LA SALLE – UNIVERSITAT RAMON LLULL
Quatre Camins 2, 08022 Barcelona (Spain)

atrilla@salle.url.edu, falias@salle.url.edu, st18187@salle.url.edu

Abstract

In the context of text processing for Text-to-Speech (TTS) synthesis, this work aims to automatically direct the expressiveness in speech through tagging the input text appropriately. Since the nature of text presents different characteristics according to whether it is domain-dependent (related to its topics) or sentiment-dependent, it is studied how these traits influence the identification of expressiveness in text.

To this end, two principal Text Classification (TC) methods are considered: a graph-based approach named the Reduced Associative Relational Network and the Maximum Entropy classifier. Their effectiveness in domain/sentiment dependent environments is evaluated. The results indicate that moving from a domain-dependent environment to a more general sentiment-dependent environment strictly results in poorer effectiveness rates, despite the sensible direct association that sentiment provides for dealing with expressiveness. Additionally, it is also evaluated how sensitive the classifiers are to a small increase of training data, yielding a slight positive influence.

Index Terms: domain classification, sentiment classification, expressive Text-to-Speech synthesis

1. Introduction

Expression is suggested to be a manner of speaking, a way of externalising feelings, attitudes and moods – conveying information about an affective state [1]. Traditionally, the Text-to-Speech (TTS) synthesis community relates expressiveness with *emotion* [2], while the Text Analysis community focuses on *sentiment* [3]. Despite many existing TTS-related publications study these problems, as far as we know a specific study on the implications concerning the expressive information that the text includes, as well as the implications of the acoustic features that convey expressiveness in speech, is still lacking. In general, the focus is on one particular aspect (either speech or text) relying on some other existing system for the missing complementary information, regardless that this already existing system may be conceived for a purpose different from the complete system. For example, [4, 5, 6] focus on the production process of expressive synthetic speech while [7, 8, 9] focus on the extraction of relevant information from text in order to direct expressiveness in speech.

In detail, [4] detects affect in text through the identification of situations that evoke common emotional responses (based on a former psychological study), and then concentrates on synthesis. In [5] and [6] the authors employ different dictionaries of affect (based on different models of affect) to extract emotional information from the lexicon, and again are focused on synthesis. On the contrary, [7] focuses on extracting linguistic features from text in order to predict its sentiment, and then synthesises speech. In [8] the authors produce a content-dependent list of emotional words to match the words in a given text, to then classify its emotion for further speech synthesis. And [9] concentrates on classifying texts pertaining to different domains (topics) and then assigns an expressive style to each domain based on a predefined expert criterion (relating each domain with a pre-defined speaking style).

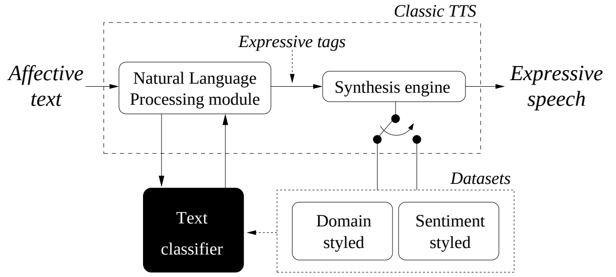

Figure 1: *Framework of a Text-to-Speech (TTS) synthesis system including a text classifier to automatically direct the expressive style in speech. Dealing with a domain-styled dataset or a sentiment-styled dataset determines the expressiveness of the TTS system, for one given language.*

This paper is focused on the automatic extraction and tagging of expressive information from text at sentence level, moving from domain to sentiment dependent characteristics, and considering its implications regarding the expressiveness of the TTS synthesis system, see Figure 1. These tags should then direct the expressive style in speech. In this study two expressiveness-enabled environments are presented: (1) a domain-styled dataset, i.e. a multidomain dataset where the domains of the texts are heuristically assigned to – and recorded with – particular expressive styles [9], and (2) a sentiment-styled dataset, i.e. a dataset where the sentiments evoked by the texts directly determine the expressive styles [10]. In domain-styled data, the lexicon pertains to a given topic, whereas in sentiment-styled data, this association may not be direct or even it may not exist at all. For example, while it is plausibly clear that words like “teacher” and “school” pertain to the same domain, namely “education”, it is unclear that these same words by themselves alone may be related with a particular sentiment. The relevance of this domain dependency to attain a good effectiveness rate in text classification is studied in environment 1. Nevertheless, dealing only with such domain-styled data sensibly limits the generalisation of expressive TTS synthesis system as it only succeeds in identifying texts pertaining to the training domains, and thus delivering speech in only the predefined domain-related expressive speaking styles. Therefore, environment 2 investigates how the text classification method performs in a more general environment, that is the sentiment-styled text, where the sentiment labels directly determine the expressive styles. Also in this work it is studied if training the classifiers with an additional small amount of text may improve the text classification effectiveness in both expressive environments. In expressive Unit Selection TTS (US-TTS), most (if not all) of the texts need to be recorded with expression [9], and the creation and labelling of a speech database is a cost to minimise. Thus, there is an interest in maintaining a restriction on the small size of the corpora because the final goal is the production of synthetic speech. In addition, real-time speech synthesis is pursued, hence, the Text Classification (TC) computer cost should also be minimised.

Section 2 describes the TC methods and corpora considered to conduct the experiments, which are described in Section 3. The obtained results are discussed in Section 4 and the paper is concluded in Section 5.

2. Text classification method

This section explores how TC performs in domain/sentiment dependent environments. To this end, two domain and sentiment corpora labelled in expression are used to to train three different principles of classification given the task at hand.

2.1. Expressiveness-styled corpora

On the one hand, the Advertising Database compiles 1350 advertisements pertaining to three different domains (topics), namely education, technology and cosmetics. Each of these domains was assigned to an expressive speaking style, happy, neutral and sensual respectively, based on an expert criterion [9].

On the other hand, the Semeval 2007 training dataset [10] compiles 1000 headlines, labelled in emotion after conducting a subjective survey, and adapted to the sentiment label set (positive/negative/neutral) through a mapping on Russell's model of affect, see [11] for further details.

Table 1: *Properties of the corpora. 5-lexicon represents the size of the lexicon of words appearing at least 5 times.*

Property	Advert. Database	Semeval 2007
Instances	1350	1000
Vocabulary	2643	3145
5-lexicon	368	212
Class-balance	0.39 (HAP) 0.38 (SEN) 0.23 (NEU)	0.55 (NEU) 0.33 (NEG) 0.12 (POS)

Note that the datasets are of comparable size and are also labelled with the same amount of categories, see Table 1. Also, for both datasets the instances correspond to sentences, i.e. one sentence per document, that is the hardest context to attain a good classification effectiveness [9]. Besides, observe the different nature of the data: while the Semeval 2007 dataset is slightly smaller, its vocabulary size is slightly greater, denoting a richer lexicon, less bound to a domain (topic) in particular. Nevertheless, the TC performance considering the nature of the data might still be influenced by the balance of the classes, but the overall results should be faithfully comparable.

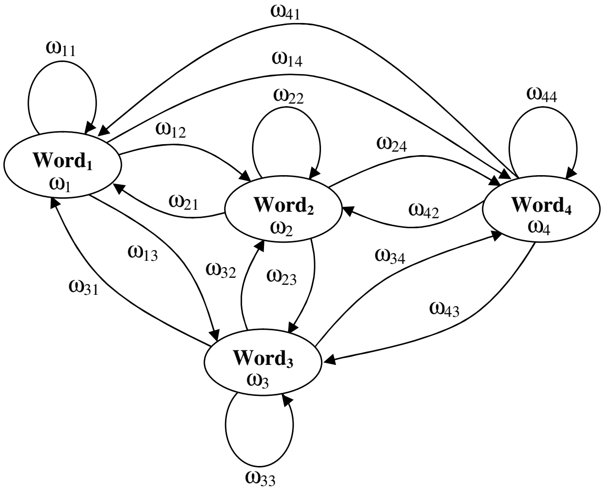

Figure 2: *Graphical representation of the ARN [9].*

2.2. Classification methods description

This section first describes the TC approaches for domain-styled text. To this purpose, the Reduced Associative Relational Network (ARN-R) is used given its conception for multidomain US-TTS synthesis and its promising effectiveness in domain-dependent TC at sentence level [9]. This method showed to perform better than Nearest Neighbour (example-based classifier), Independent Component Analysis (unsupervised text classifier) and n-grams at character-level (inductive generative classifier) in [9]. Support Vector Machines (SVM), which are some of the most successful techniques for conducting TC, were determined to fail at classifying the Advertising Database due to its small size [9]. Thus, SVM are excluded from the experiments given the size of the datasets.

However, in order to extend the performance comparison of the ARN-R, it is compared against a Maximum Entropy (MaxEnt) classifier. On the one hand, due to its functional form, that is different from the rest of the already compared classifiers (inductive discriminative classifier). And on the other hand, due to its use with success in many sentiment analysis tasks [12], with a performance comparable to the SVM. Hence, its consideration is found necessary to obtain a good baseline of the current state of the art.

Then the work continues on describing the TC approaches for sentiment-styled text. The ARN-R and MaxEnt classifiers are compared, now on the sentiment environment, and additionally considering a Nearest Emotional Centroid (NEC) classifier (centroid-based classifier with expert knowledge) [11]. The NEC is regarded to be a good baseline because it intuitively captures the expert knowledge related to the field of emotion.

2.2.1. Reduced Associative Relational Network

The ARN-R builds a graph of words for each training category by associating each word to a node and each collocation (ordered co-occurrence of two adjacent words) to a directed link, like in Figure 2. The weights associated to each term (node or link) are computed from training corpora and weighted according to a term weighting method.

On testing, the ARN-R first builds a similar graph with the text to test. Afterwards, it vectorises its weighted terms defining the dimensions of a Vector Space Model (VSM). In this space, a vector for each category is projected, maintaining the weights of each term coinciding with the dimensions of the VSM. In the end, the ARN-R makes use of a similarity measure (e.g., the cosine distance) and assigns the label corresponding to the most similar categorical representation (hard classification).

Regarding the term weighting methods, the ARN-R considers an unsupervised method named the Inverse Term Frequency (ITF), which yields better results than simple term frequencies, see [9]. Moreover, a supervised weighting method, named the Relevance Factor (RF), is also considered. The RF yields better results than traditional unsupervised term weighting methods in TC, see [13]. Equations (1) and (2) show the ITF and the RF respectively.

$$ITF_t = \log\left(\frac{\sum_{t' \in T} tf_{t'}}{tf_t}\right) \qquad (1)$$

$$RF_{t,C} = \log\left(1 + tf_t\right)\ \log_2\left(2 + \frac{tf_{t,C}}{\max(1, \sum tf_{t,\bar{C}})}\right) \qquad (2)$$

where t represents the term, tf represents its Term Frequency, T represents the vocabulary (total number of different terms) and C represents the positive category (likewise $\bar{C}$ represents the negative category).

While the ITF intends to weight the local contribution of a given term, the RF intends to weight the contribution of a term considered to pertain to a given category regarding the rest of categories.

2.2.2. *Maximum Entropy*

Maximum entropy modelling is a framework for integrating information from many heterogeneous information sources for classification. MaxEnt models are first given a set of constraints that are justified by the available data, and then compute the model with maximum entropy of all the models that satisfy the constraints. The MaxEnt model is motivated by the desire to preserve as much uncertainty as possible, avoiding to infer anything beyond the data, see [14] for further details. The categories are modelled with the exponential form shown in Equation (3).

$$P(C|t) = \frac{1}{Z}\ \exp\left(\sum_i \lambda_{i,C}\ F_{i,C}(t, C)\right) \qquad (3)$$

where the Z above is a normalisation factor in order to define a probability distribution. The feature/category functions $F_{i,C}$ deal directly with the term binary features of presence or absence without assuming any relationship among them. The parameters $\lambda_{i,C}$ are set to maximise the entropy of the induced distribution. The expected values of the feature/class functions have to be equal to the evidence shown in the training data [12].

2.2.3. *Nearest Emotional Centroid*

The NEC first represents text in a space defined by emotional dimensions (basic properties of affective states according to expert knowledge: valence, activation and control), named the circumplex. Then, it computes the sentiment centroids on training data (i.e. their vector sum). Finally, it performs comparisons with a minimum-distance (to the centroids) criterion, see [11] for further information. The NEC approach is considered to be a suitable baseline to deal with such texts of subjective nature (the sentiment of affective states) for incorporating offline expert knowledge.

3. Experiments

Firstly, the classifiers are submitted to experimentation on both datasets, the Advertising Database and the Semeval 2007 dataset, in order to study how the considered classification methods perform on domain versus sentiment-styled texts respectively. To that effect, several configurations are considered: in the term weighting method (ITF or RF) and in the consideration of collocations.

Moreover, it is studied if slightly augmenting the size of the training data could be of help to improve the effectiveness rates. According to the premise of small-sized datasets, a 10% data increase is appended to the Advertising Database (the largest dataset), and a 25% data increase is appended to the Semeval 2007 dataset (the smallest dataset). The different size in these extensions intends to make the corpora more similar with regard to their size.

The experiments are compared using the macroaveraged F_1^M effectiveness measure [15], estimated with a 10-fold Cross Validation method. ANOVA tests at the 0.05 confidence level are used for evaluating statistical significance.

4. Results and discussion

The results are shown in Table 2. A preliminary experiment with the ARN-R with plain term frequencies (no term weighting method applied) yielded an average effectiveness rate of 58.69% for the Advertising Database, and 46.73% for the Semeval 2007 dataset. This experiment intended to highlight the importance of the term weighting method in the TC task, given that any of the presented methods with weighted terms has yielded better results than plain term frequencies (8.54% better on average).

Regarding the relevance of domain dependence to attain a good effectiveness rate in TC, as suggested in Section 1, the obtained results indicate that domain dependence contributes positively toward a good effectiveness rate. It can be observed that the ARN-R and MaxEnt methods perform equivalently (with no significant difference) in the two environments despite their different learning paradigm. They yield almost a 21% of average superior performance for the domain-styled environment (the Advertising Database) regarding the non domain dependent, but generalisable, sentiment-styled environment (the Semeval 2007 dataset). These high effectiveness rates are attributed to the relation between the lexicon and the domain, i.e. the typical case in TC [15]. When the TC method intends to match an unknown text to any of the learnt domain-dependent words, it performs better when there are many words alike (note that the Advertising Database contains 7.18% more frequent words than the Semeval 2007 dataset, see “5-lexicon” field in Table 1).

Nevertheless, a significant difference is observed among the features used for classification. In the domain dependent environment, the ITF has performed significantly better than RF, about 20% better without collocations, and about 11% better with collocations. The fact that the RF performs better with collocations responds to the theory that the RF raises the presence of singularities in the data [9], and such singularities are more easily found among collocations than among words alone. In the sentiment dependent environment, though, all classification methods have performed similarly regardless of their principle

Table 2: *Results evaluating the F_1^M effectiveness rate of several text classification methods on different environments (mean ± std).*

Classification method	**Advertising Database**		**Semeval 2007 corpus**	
	Training	Extension	Training	Extension
ARN-R ITF	79.73% ± 3.69	79.81% ± 3.13	51.45% ± 4.71	51.81% ± 4.94
ARN-R ITF + Col.	79.15% ± 4.36	79.40% ± 3.08	52.27% ± 6.02	54.09% ± 5.45
ARN-R RF	59.36% ± 4.30	60.06% ± 4.28	53.96% ± 3.71	55.93% ± 4.09
ARN-R RF + Col.	68.56% ± 3.75	70.45% ± 3.73	55.78% ± 2.86	56.18% ± 3.35
MaxEnt	79.95% ± 3.47	80.36% ± 2.59	51.74% ± 7.95	53.04% ± 7.99
MaxEnt + Col.	76.30% ± 3.19	77.56% ± 2.33	52.18% ± 6.94	53.89% ± 6.33
NEC	–	–	49.04% ± 5.33	48.96% ± 5.24

of classification or term weighting strategy. All methods have yielded better results than the baseline NEC, although no statistically significant difference is observed among them. These results show the difficulty of classifying text pertaining to a more general environment, where the lexicon may not be related with a particular domain.

Regarding the training of the classifiers with an additional small amount of text, the results show a positive tendency in all cases (no particular behaviour depending on the features nor the classification principle), but the improvements are not statistically significant. The maximum increment observed is close to 2%. This increment is expected to grow as the size of the extension is enlarged, but then the cost of recording a large dataset should be faced (the trade-off needs to be studied in detail). In summary, the overall results indicate that while it is positive to slightly extend the training dataset, the final effectiveness rates are insignificantly increased.

5. Conclusions

This work intends to contribute to the generalisation of sentence-level TC for expressive TTS synthesis. In this sense, the impact of moving from domain-dependent to sentiment-dependent expressiveness in text is analysed, because the latter has a more sensible direct association with expressiveness. To that end, the ARN-R and MaxEnt TC systems are evaluated. They yield an equivalent effectiveness performance in each of the studied domain-dependent and sentiment-dependent environments. However, for domain-dependent data results are significantly better. This may be due to the strong relation between the data domain and its lexicon. The results delivered by both systems for sentiment data outperform the baseline NEC system, and scoring better in this environment already represents an improvement with respect to the expressiveness generalisation purpose.

The ARN-R presents more flexibility than MaxEnt to use strategies related to the text processing field, such as the term weighting schemes. In this sense, this work has evaluated the impact of unsupervised (ITF) and supervised (RF) term weighting functions, obtaining slightly better results (non-significant) with the supervised method. In the future work, in addition to considering more term weighting aspects, it is expected to extend the exploitation of the similarity of texts in the ARN-R with graph-based measures, like graph distances such as the Pattern Length [9]. These measures are expected to strengthen the need of a graph-based structure to better model the behaviour of sentiment-styled text.

6. References

[1] M. Tatham and K. Morton, *Expression in Speech: Analysis and Synthesis*. New York, NY, USA: Oxford University Press, Inc., 2004.

[2] P. Taylor, *Text-to-Speech Synthesis*. Cambridge, UK: Cambridge University Press, 2009.

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[4] A. R. F. Rebordao, M. A. M. Shaikh, K. Hirose, and N. Minematsu, "How to Improve TTS Systems for Emotional Expressivity," in *Proc. of Interspeech'09*, Brighton, UK, Sep. 2009, pp. 524–527.

[5] Z. Wu, H. M. Meng, H. Yang, and L. Cai, "Modeling the Expressivity of Input Text Semantics for Chinese Text-to-Speech Synthesis in a Spoken Dialog System," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 8, pp. 1567–1576, Nov. 2009.

[6] G. O. Hofer, K. Richmond, and R. A. J. Clark, "Informed Blending of Databases for Emotional Speech Synthesis," in *Proc. Interspeech*, Sep. 2005.

[7] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proc. of HLT'05*. Morristown, NJ, USA: ACL, 2005, pp. 579–586.

[8] V. Francisco and R. Hervás, "EmoTag: Automated Mark Up of Affective Information in Texts," in *Proc. of the Doctoral Consortium in EUROLAN 2007 Summer School*, C. Forascu, O. Postolache, G. Puscasu, and C. Vertan, Eds., Iasi, Romania, Jul.–Aug. 2007, pp. 5–12.

[9] F. Alías, X. Sevillano, J. C. Socoró, and X. Gonzalvo, "Towards High-Quality Next-Generation Text-to-Speech Synthesis: A Multidomain Approach by Automatic Domain Classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 7, pp. 1340–1354, Sep. 2008.

[10] C. Strapparava and R. Mihalcea, "SemEval-2007 Task 14: Affective Text," in *Proc. of SemEval'07*, Prague, Czech Republic, Jun. 2007.

[11] A. Trilla and F. Alías, "Sentiment classification in English from sentence-level annotations of emotions regarding models of affect," in *Proc. of Interspeech'09*, Brighton, UK, Sep. 2009, pp. 516–519.

[12] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proc. of EMNLP'02*, Philadelphia, PA, USA, Jul. 2002, pp. 79–86.

[13] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization," *IEEE T. Pattern. Anal.*, vol. 31, no. 4, pp. 721–735, Apr. 2009.

[14] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: The MIT Press, 1999.

[15] F. Sebastiani and C. N. D. Ricerche, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, pp. 1–47, 2002.