Diploma d'Estudis Avançats

Natural Language Processing techniques applied to speech technologies

(supervised work)

Author : Alexandre Trilla Castelló

Advisor : Dr. Francesc Alías Pujol

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

2010

Abstract

This dissertation relates a first research stage in the pursuit of the Ph.D. degree in "Information and communication Technologies and its Management" at La Salle – Universitat Ramon Llull.

The dissertation presents an overview of the various approaches to Text-based Sentiment Prediction in order to reveal their effectiveness in the tripartite sentiment recognition task, i.e., the identification of positive, negative and neutral orientations in text. It discusses the application of the bases that support these diverse proposals, from the feature extraction stage to the classification phase, contrasting the contributions that each method yields through a set of experiments with the Semeval 2007 dataset and the Fifty Word Fiction corpus.

Contents

Contents i

List of Tables iii

List of Figures iv

1 Supervised work 1

1.1 Introduction . . . 1
1.2 Sentiment analysis task . . . 4
1.3 EmoLib: sentiment identification from text . . . 6
1.4 Lexical affective features . . . 7
1.4.1 Dimensional theories of emotion . . . 7
1.4.2 Extracting affect features from text . . . 9
1.4.3 Similarity measures . . . 11
1.5 Principle of classification . . . 12
1.5.1 Heuristic classification . . . 12
1.5.2 Data-driven classification . . . 12
1.6 Hierarchical and risk-assessed strategies . . . 16
1.6.1 Minimum-risk: posterior weighting strategy . . . 16
1.6.2 Decision levels: hierarchical strategy . . . 19
1.7 Experiments . . . 20
1.7.1 Sentiment datasets . . . 20
1.7.2 TSP schemes . . . 25
1.8 Results and discussion . . . 27
1.8.1 Dataset comparison . . . 29
1.8.2 Features comparison . . . 29
1.8.3 Classifier comparison . . . 30
1.8.4 Strategy comparison . . . 34

2 Conclusions and future work **35**

3 Contributions **38**
3.1 Scientific publications 38
3.2 Associated research projects 39

Bibliography **41**

List of Tables

1.1 Example words pertaining to the ANEW dictionary of affect. For indexing purposes, the actual entries of the dictionary are the stems of the words. Note the emotional dependency on the POS tag. 9
1.2 Example relations between situations and emotional categories according to [Shaver et al., 1987]. 12
1.3 Relation between emotional dimensions and categories for the system proposed in [García and Alías, 2008]. 13
1.4 Tagging examples of the Semeval 2007 dataset. "A" stands for Anger, "D" for Disgust, "F" for Fear, "J" for Joy, "Sa" for Sadness and "Su" for Surprise. 21
1.5 Distribution of emotions in the Semeval 2007 dataset according to the considered models of affect. 23
1.6 Adequacy of the considered models of affect for mapping the Semeval 2007 dataset. 24
1.7 Properties of the sentiment datasets. 4-lexicon represents the lexicon of words appearing at least 4 times. 25
1.8 Results from the comparative study. Terms are considered to be single words. The two best results for each corpus are printed in boldface, one for emotional dimensions and the other for textual features. 28
1.9 List of the ten most frequent words and tuples and their distribution in the Semeval 2007 dataset, expressed in terms of conditional probability percentage (approximated). 31
1.10 List of the ten most frequent words and tuples and their distribution in the FWF corpus, expressed in terms of conditional probability percentage (approximated). 32

List of Figures

1.1 Block diagram of a TTS synthesis system including a sentiment text classifier following the approach introduced in [Alías et al., 2008] for conducting multidomain TTS synthesis. 3
1.2 Taxonomy of emotion [Shaver et al., 1987]. Shaver and colleagues put "surprise" in brackets because some authors don't consider it a prototypical emotion. This convention is maintained in the description of the taxonomy. 5
1.3 EmoLib processing framework diagram. 6
1.4 The structure and sequentiality of text in the Associative Relational Network [Alías et al., 2008]. 14
1.5 Taxonomy of emotion considered in this work. Although the main purpose is to attain sentiment classification, emotions are also shown here for conceptual clarity. 20
1.6 Distribution of basic emotion categories in the circumplex according to Russell's model of affect. 22
1.7 Distribution of basic emotion categories in the circumplex according to Whissell's dictionary of affect. 23

Chapter 1

Supervised work

This chapter describes the supervised research work carried out up to present. Firstly, the environment that frames this work is presented, describing the research group within it is conducted and its integration with its research areas.

Secondly, the concepts learnt in the training stage are put into practice. The principal theoretical topics of the research line are introduced, a set of experiments are conducted and some tentative contributions are proposed as a means of improvement.

Finally, the discussion motivated by the yielded results is produced to conclude the research work of the dissertation.

1.1 Introduction

The human speech communication process can be thought of as comprising two channels: the words themselves and the style in which they are spoken. Each of these channels carries information [Eide et al., 2004]. Relating to the second channel, it is feasible to associate its characteristics with affective states. This connection may be attained through speaking with an expressive style [Hofer et al., 2005, Alías et al., 2008]. The present work explores the automatic extraction of affective information from text for further use in a Text-to-Speech (TTS) synthesis scenario. As a first step towards this expressive TTS synthesis the dissertation formalises its objective into a *sentiment analysis* task, i.e., the identification of positive, negative and neutral stances in text. This goal is aimed at the first level of the hierarchy of emotion [Shaver et al., 1987].

Sentiment analysis is a topic that has gained interest and popularity in time. Presently, with the increasing demand of a more natural Human-Computer Interaction (HCI), the sentiment space is one of the key aspects to understand the implicit channel of communication [Cowie et al., 2001], which transmits non-verbal messages along with the explicit verbal messages (say the objective information). This "reading between the lines" has traditionally been tackled by psychology, trying to build an emotional knowledge base (related to sentiment) to deal with these recognition/classification aspects.

This dissertation frames the sentiment analysis task in a textual environment. Therefore, the limited access to clear information (texts can be confusing) and/or the lack of alternative modes (that may help to resolve ambiguities) makes it an especially challenging task. In this setting, all the words extracted from the text are of vital interest to predict the expressed sentiment, as the *word* by itself is plausibly conceived as the "smallest meaningful affective unit" [Batliner et al., 2009]. What is more, the polarity of a word, i.e., its semantic orientation, may be dependent on its context, and therefore the sentiment conveyed in a given text (represented as a collection of words) is the result from the interaction among the polarities of the words within. For the TTS interests pursued in this work, the given text is framed into sentences, as the sentence is the smallest expressive unit considered. Then, each sentence tagged with sentiment is related to a particular speaking style in order to be synthesised accordingly [Alías et al., 2008].

The identification of affective information with discrete sentiment categories (related to the speaking style), is in concordance with the TTS design developed by the Speech Research Group at the university [Monzo et al., 2008]. Thus, the automatic classification of input text is of great interest as a value-add of the speech synthesis engine, in contrast to including explicit text tagging, which is almost impossible to get for general purpose applications. See Figure 1.1 for a descriptive representation of the intended system. Moreover, this system could also be attached to the processing chain in an Automatic Speech Recognition engine in order to deal with the semantics of the recognised utterances, identifying its semantic orientation, and thus helping to improve its accuracy.

The analysis of sentiment entails dealing with the emotional, and hence subjective, character of text. To this end, the research community has proposed several methods, some of which are based on heuristic rules (Knowledge Engineering), like SenseNet [Shaikh et al., 2008] and EmoLib [García and Alías, 2008], and some data-

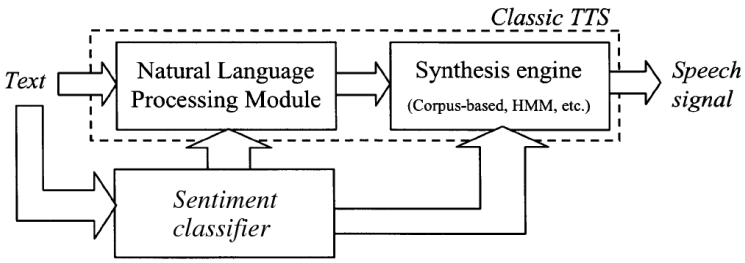

Figure 1.1: Block diagram of a TTS synthesis system including a sentiment text classifier following the approach introduced in [Alías et al., 2008] for conducting multidomain TTS synthesis.

driven, like EmoTag [Francisco and Hervás, 2007] or [Wilson et al., 2009, Strapparava and Mihalcea, 2008, Alm et al., 2005]. However, the rise of the widespread availability to researchers of organised data labelled with sentiment has contributed to a large shift in direction toward data-driven approaches [Pang and Lee, 2008]. Some of these systems also introduce the final objective pursued in this work, the expressive TTS synthesis (see [Rebordao et al., 2009] and [Francisco et al., 2007]), along with the works of [Hofer et al., 2005] and [Alías et al., 2008].

In this data-driven line, many proposals make use of Support Vector Machines (SVM), see [Wilson et al., 2009, Li and Zong, 2008, Abbasi et al., 2008, Pang and Lee, 2005, Pang et al., 2002]. Other works also include boosting, memory-based and rule-based learning [Wilson et al., 2009], genetic algorithms [Abbasi et al., 2008] and Naive Bayes plus Maximum Entropy [Pang et al., 2002]. Besides, in the revised bibliography, there is a different approach that builds a graphical structure and then exploits its properties with a modified PageRank algorithm [Cruz et al., 2009].

With regard to the features used, some systems make use of plain linguistic features, e.g, presence/frequency of certain Part-Of-Speech tag sequences, like [Wilson et al., 2009, Shaikh et al., 2008, Li and Zong, 2008, Abbasi et al., 2008, Pang et al., 2002]. An alternative approach may also consider the use of Information Retrieval (IR) measures [Pang and Lee, 2005].

This work is based on EmoLib [García and Alías, 2008], a library built entirely upon vocabulary expert knowledge to tag the emotion of input

text. In this dissertation it is discussed how this system can be enhanced through the knowledge automatically acquired from more complex linguistic structures: sentence-level annotations of emotions considering models of affect (for dealing with their sentiment), thus biasing the modus operandi of EmoLib towards a data-driven approach. This corpus labelling approach (annotations of emotions at sentence-level) is believed to increase the effectiveness of the system compared to positive/negative valence annotation alone [Strapparava and Mihalcea, 2007] (informal experiments point towards this belief). The resulting scheme should stand as a first step towards automatic expressive sounding TTS synthesis.

All in all, the determination of the current research is to better understand the nature of the sentiment analysis problem, provide salient features denoting these affective aspects at sentence-level, and accurately design a classifier in consonance with these aspects altogether in order to reach a maximum classification effectiveness. To that effect, Section 1.2 presents the task attacked in this work and Section 1.3 introduces the framework of the system developed to this end. Then, Section 1.4 describes the set of salient features used to represent text and the assessment of their similarity, Section 1.5 describes the classification strategies used, Section 1.6 applies different architectures to the classification schemes as a means of refinement, Section 1.7 presents the experiments and Section 1.8 discusses the obtained results.

1.2 Sentiment analysis task

The previous work of the research group in this field is compiled in [Alías et al., 2008, García and Alías, 2008]. While [Alías et al., 2008] tackled the problem of text classification according to its topic (then relating the topic to a particular speaking style), [García and Alías, 2008] presented a more general task addressing the six prototypical emotions, i.e., the identification of angry, fearful, sad, neutral, happy and surprising stances in text. This task is adequate for the TTS synthesis technology developed at the Speech Research group at the university (see [Monzo et al., 2008] for the TTS system description). Nevertheless, as it can be observed in [García and Alías, 2008], some emotional labels, e.g, "surprise" and "sorrow", cannot be identified, at least with the strategy they propose, that is the representation of emotions in a circumplex and the use of a heuristic rules classifier. If one of the covered classes cannot be predicted, then its precision rate cannot be computed (see

Section 1.8), and this is a problem for comparing its performance with other strategies.

As an improvement, it may be sensible to consider different classification strategies, and/or less classes, despite the latter is more a simplification rather than an improvement. Nevertheless, for the particular case of considering sentiments, i.e., negative, positive and neutral cover classes, and also through considering Shaver's hierarchy of emotion, see Figure 1.2, these two worlds (sentiment and emotion) might be related. Therefore, the task of sentiment classification could be regarded as a previous step to attain the emotion classification needed for the TTS synthesis engine.

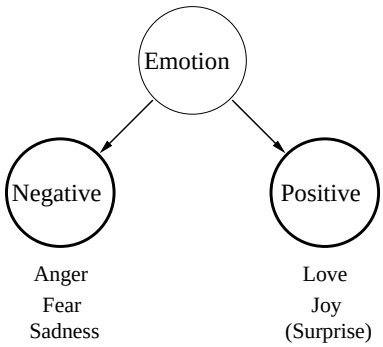

Figure 1.2: Taxonomy of emotion [Shaver et al., 1987]. Shaver and colleagues put "surprise" in brackets because some authors don't consider it a prototypical emotion. This convention is maintained in the description of the taxonomy.

Perhaps sentiments alone cannot effectively discriminate emotions at the spoken speech level. In particular, [Schröder, 2004b] shows typical prosodic (stylistic) speech features, e.g., F_0 mean, F_0 range, tempo or loudness, for a set of prototypical emotions (joy, sadness, anger, fear, surprise and boredom). If these parameters for the emotional classes are intended to be represented with one sentiment identifier, a compromise cannot be reached (e.g., incompatible specifications). This is due to the fact that the hierarchical classification of emotion is "only" compliant with the valence property of emotion, regardless of the activation property (see Section 1.4.1 for the approach with emotional dimensions). Thus, a positive (or negative) sentiment identifier cannot differentiate between passive or active positive (or negative) emotions. However, this well-formed argument is still dependent on the

speech corpus design considerations. As an example, the FAU Aibo Emotion Corpus used for the Interspeech Emotion Challenge [Schuller et al., 2009] is labelled with a mixture of sentiments and emotions: emphatic, angry, neutral, positive and rest (that includes bored, helpless and surprised).

To sum up, despite sentiments have their limitations to model emotional speech style on their own, they may be considered a first step to attain a complete expressive model for TTS synthesis. And most importantly, their consideration possibly avoids the problem of missing some emotional cover classes. This issues are treated in the experimentation part (Section 1.7).

Finally, given that the classification problem aimed at this work is centred on sentiments, it may be sensible to name it the Text Sentiment Prediction (TSP) problem. This nomenclature is used from now on to refer to this classification problem.

1.3 EmoLib: sentiment identification from text

The original EmoLib architecture, described in [García and Alías, 2008], firstly extracts relevant features from text and then it applies a classifier to assign the most appropriate emotional tag to the text being analysed, as is shown in Figure 1.3. In this work, though, since the interest is focused on sentiment analysis, the emotion labels are grouped into sentiment labels according to the hierarchy of emotion shown in Section 1.2. The different modules that build EmoLib are described hereunder.

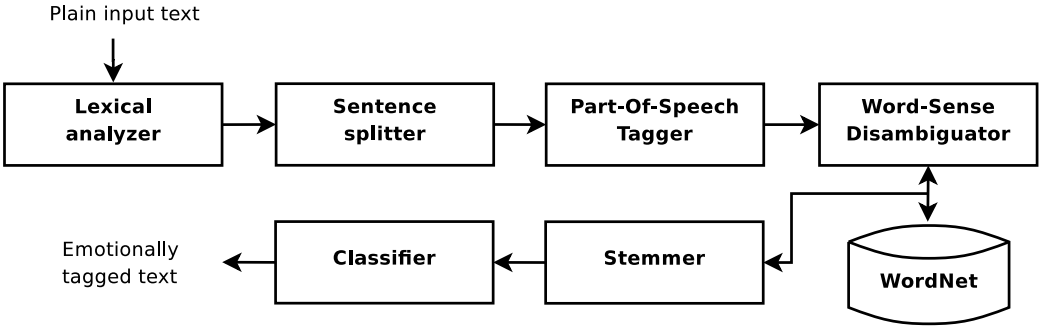

Figure 1.3: EmoLib processing framework diagram.

Lexical analyser: Converts the plain input text into an output token stream. It spots the possible affective containers (nouns, verbs, etc.) filtering out the rest of affectively irrelevant particles (prepositions, articles, etc.), also known as "stop words". This module is produced with the JavaCC[1] parser generator.

Sentence splitter: Determines the boundaries that limit sentences and splits the input text into sentences with a binary decision tree inspired in [Reichel and Pfitzinger, 2006].

Part-Of-Speech (POS) Tagger: Determines the function of nouns, verbs and adjectives (affect containers) in the sentence using the Stanford log-linear POS tagger [Toutanova and Manning, 2000].

Word-Sense Disambiguator: Determines the meaning of nouns according to the context. Additionally provides a set of synonyms for the resulting sense. In this module the WordNet database [Fellbaum, 1998] is of use.

Stemmer: Removes the inflection of words for indexing purposes using the Porter stemming algorithm [Porter, 1980]. Related words should map to the same stem, base or root form.

Classifier: Classification scheme to accomplish the affective classification of input text.

Having this processing pipeline yielding a host of linguistic features from text, what follows next is the procedure to extract salient features from text and the definition of adequate strategies to solve the TSP problem successfully.

1.4 Lexical affective features

Considering words as the smallest units containing affect, this section explores their representation and similarity evaluation in different feature spaces.

[1] https://javacc.dev.java.net/

1.4.1 Dimensional theories of emotion

Many psychological studies reported in the literature lead to an unified dimensional emotional space to represent affective concepts (e.g., the sentiment categories) and link the dimensional ratings to the application area of interest – see [Schröder, 2004a] for a formulation of emotional prosody rules or [Li and Ren, 2009] for emotional orientation prediction from text.

Emotional dimensions are though a simplified description of basic properties of emotional states [Schröder et al., 2001]. While they do not capture all the relevant aspects of an emotional state, they provide a taxonomy allowing simple distance measures between emotion categories to be used to contrast these basic properties. This approach has historically been embraced for data-driven research activities [Cowie et al., 2001] and recently it has been adopted by the W3C with the EmotionML specification [Baggia et al., 2008].

In the literature, one of the most popular emotion evaluation spaces is the *circumplex*: a bidimensional space that represents the valence (positive/negative evaluation) and the activation (stimulation of activity) of emotions. Regarding the specific location of the emotional categories in the circumplex, this approach has though some slight differences according to the considerations taken by their authors, thus defining different models of affect. For instance: Russell's affective model [Russell, 1980], Scherer's model [Scherer, 1984], Plutchik's model [Plutchik, 1980], Watson and Tellegen's model [Watson and Tellegen, 1985], and Whissell's dictionary of affect [Whissell, 1989]. Russell's model of affect appears in [Russell et al., 1989] as a reference circumplex through a figure with a setting of points representing the emotions. The numerical data has been obtained from the relative position of the points in the canvas. Whissell's model of affect, used in [Hofer et al., 2005], appears in [Cowie et al., 2001] contrasted with the completely different approach to emotional dimensions that Plutchik proposed [Plutchik, 1980], arranged in an "emotion wheel" instead of a circumplex. As it can be seen in the extensive table provided in the article, the emotional values have some significant differences with Russell's. These differences imply a different location of the basic emotions in the circumplex, which in its turn it is sensible to believe that the classification approach will be more or less biased. And finally, an adapted Scherer's model of affect appears in [Généreux and Evans, 2006] as a reference circumplex for the binary classification experiments presented. Some later works [Mehrabian, 1995, Bradley and Lang, 1999] also intend to measure a comple-

mentary emotional feature/dimension for a given environment, the control or power (domination that exerts on the subject), in order to grasp the finest distinctions between emotions. These three characteristics of emotion are nearly independent [Mehrabian, 1995].

Empirical applications of the dimensional model using words as affective stimuli have been successfully used in many studies, see [Stevenson et al., 2007]. Moreover, affective words substantially contribute in indicating the sentiment of a sentence or document (averaging strategy) [Kim and Myaeng, 2007]. Preliminary experiments indicate that as long as this compositional approach contemplates linguistic facts such as context-dependent semantics or negations, the resulting predictions are more successful.

The knowledge that relates a lexicon to a set of emotional dimensions is compiled in a *dictionary of affect*, e.g., the Affective Norms for English Words (ANEW) [Bradley and Lang, 1999], the Dictionary of Affect in Language (DAL) [Whissell, 2008] and some Lists of Emotional Words (LEW) based on discriminant lexicons [Osherenko, 2008, Francisco and Gervás, 2006]. There is some disagreement among these models, though, that is attributed to dictionary design purposes and the personality, mood, social background and context situation of the evaluators. For example, the ANEW dictionary of affect was created specifically by a psychological study measuring the associations between words and human emotions, and contains 1035 words scored for valence, activation and control with the Self Assessment Manikin graphical tool [Bradley and Lang, 1999], see Table 1.1 for some words pertaining to ANEW shown as an example of a dictionary of affect. Conversely, the DAL was designed to measure the emotional meaning of words and texts, and contains a list of 8742 words rated by people for their activation, evaluation and imagery (a similar set of emotional dimensions).

1.4.2 Extracting affect features from text

Textual features are generally defined by *terms*. Terms can be set at various granularity levels, such as words, co-occurrences, phrases, sentences or any other semantic and/or syntactic units used to identify the contents of a text. Thus, terms interact dependently. For indexing purposes, they may be assembled in a LEW and used as TSP tools through lexical affinity methods [Osherenko, 2008, Valitutti, 2004].

Explicit textual features imply that the term (or its reasonable synonym)

Word	Valence	Activation	Control	POS tag
aggress	1.96	6.94	3.69	adjective
astonish	5.52	5.96	4.26	adjective
capabl	7.52	5.92	6.70	adjective
chair	5.03	3.48	5.57	noun
destroy	2.09	6.29	3.97	verb
love	8.50	7.46	5.79	noun
love	7.99	6.43	5.79	verb

Table 1.1: Example words pertaining to the ANEW dictionary of affect. For indexing purposes, the actual entries of the dictionary are the stems of the words. Note the emotional dependency on the POS tag.

appears in the text [Liu, 2010]. Term presence alone (regardless of its number of occurrences), though, misses information of discriminant value. Therefore, term frequencies are of use to capture the "strength of evidence" that is needed for the Text Categorisation (TC) task, yet simple text counts can give some sort of indication of style or authorship [Manning and Schütze, 1999]. But beyond simply scoring the number of appearances of terms, the Term Weighting (TW) scheme stands as an important step to improve the effectiveness of TC [Lan et al., 2009]. These weights (usually borrowed from the traditional Information Retrieval (IR) field) measure the discriminating power of terms and denote how much they contribute to TC. Traditional TW is prone to use *unsupervised* metrics, i.e., the ones that do not account for the category labels of the training examples, such as the Term Frequency Inverse Document Frequency (TF IDF), that raises a term t directly to its frequency in the corpus c and inversely to the number of documents d (in the corpus) where it appears, see Eq. (1.1), or the Inverse Term Frequency (ITF) [Alías et al., 2008], that is a local (sentence-level) approximation of IDF, see Eq. (1.2). Recent TW tends to use *supervised* metrics that do consider these class labels, such as the Relevance Factor (RF) [Lan et al., 2009], that raises a term according to its greater concentration (higher frequency) in the category of interest *cat* than in the rest of categories, see Eq. (1.3). The latter TW method, normally also weighted with the frequency of the term, is reported to achieve the best performance confirmed by experimental evidence on cross-method comparisons (feature robustness), cross-classifier (learning procedure robustness) and cross-corpus validation (domain trans-

fer robustness).

$$tfidf(t,c) = \#(t,c)\ \log \frac{\#(d,c)}{\#(t,d,c)} \tag{1.1}$$

$$itf(t,d) = \log \frac{\#(terms,d)}{\#(t,d)} \tag{1.2}$$

$$rf(t,cat) = \log\left(2 + \frac{\#(t,cat)}{max(1,\#(t,!cat)}\right) \tag{1.3}$$

In the above equations, t represents a term, c the corpus, d the document, cat the category, $!cat$ any category different from cat, and $\#(\cdot)$ computes the number of times $(\cdot)$ appears in the available data.

Other approaches to feature extraction contemplate the implicit characteristics in text, i.e., not directly observable aspects. Some of these traits are the *stylistic* features like the information about the structure and sequentiality of text [Alías et al., 2008], including function words in addition to content words, punctuation marks and their ordered co-occurrences in text. The consideration of words in a context to train the classifiers reduces some of the disadvantages associated to simple keyword spotting, because the context-dependent polarity of words may appear to be quite different from the word's prior polarity [Wilson et al., 2009].

Additionally, POS tag frequencies and *patterns* enable considering abstract representations of text, similar to the grammatical patterns from linguistic studies used in [Osherenko, 2008]. Words have probabilistic distributions wrt neighbouring syntactically related words [Mohammad and Hirst, 2005]. These distributions may be modelled with n-gram patterns with varying levels of lexical instantiation[2] (patterns of words and POS-tags together) [Murray and Carenini, 2009]. In the latter cite, this technique is proposed as a first-level polarity classification to permit the identification of *subjective* clauses in text, as they are believed to contain most (if not all) the affective content conveyed.

Finally, these patterns could also be used to describe *emotional events* [Shaver et al., 1987], and through the cognitive theory of emotions their valenced reactions would be assessed for classification [Shaikh et al., 2008]. This approach implies hard-coding rules in contrast to the former corpus-driven statistical-based proposals.

[2]Each unit of the pattern can be either a word or the word's POS.

1.4.3 Similarity measures

The similarity measures must be consistent with the nature of the different feature spaces. For TW features, for example, the weights of the terms are vectorised and represented in a n-dimensional Vector Space Model (VSM) [Salton et al., 1975]. This VSM of weighted terms interprets each vector instance as a direction (no interpretation of the norm as emotion strength like in the circumplex). Consequently, to account for these differences, the *distributional* type of measures are of use and thus analysed hereunder.

Distributional measures assess the similarity measure without considering further relations other than the distances in the feature space, i.e., the difference of similarity measures without considering any knowledge-network. Within this type of distance measures, some are *compositional*, like the Euclidean, which is suitable for the circumplex, and some are *non-compositional*, like the cosine, adequate for the VSM.

As it may be intuited from the discourse above, there also exists another type of evaluation measures, the *ontology/network-based* measures, which do employ a graph-based structure to relate terms with a knowledge rich criterion [Mohammad and Hirst, 2005], or a pattern length criterion [Alías et al., 2008], for example.

1.5 Principle of classification

This section addresses different classification methodologies and their contribution to sentiment prediction from text.

1.5.1 Heuristic classification

This section contemplates the classification strategies that are somehow based on expert knowledge. This knowledge is usually exploited through a set of rules applied on a certain environment. For example, the compilation of emotionally eliciting situations in [Shaver et al., 1987] (the aforementioned "emotional events"). This environment relates how determined situations evoke common emotions. See Table 1.2 for some examples.

Alternatively, another example of heuristic rules is [García and Alías, 2008], which works with three emotional dimensions and defines a set of intervals to delimit six emotions, see Table 1.3.

Category	**Emotionally eliciting situation**
Fear	Threat of harm or death
Sadness	Rejection, exclusion, disapproval
Anger	Aggressive, threatening gestures
Joy	Receiving esteem, respect, praise
Love	Wanting the best for other, etc.

Table 1.2: Example relations between situations and emotional categories according to [Shaver et al., 1987].

Category	**Valence**	**Activation**	**Control**
Surprise	$\geq$8.5	$\geq$6.35	$\geq$6.5
Happiness	$\geq$6.445	$\geq$5.86	$\geq$5
Sadness	$\leq$3	$\leq$4.575	$>$1.5 and $\leq$3.75
Anger	$\leq$3.25	$\geq$6.25	$>$3.5 and $\leq$4.5
Fear	$\leq$3	$\leq$7.5	$\leq$3.75
Neutral	Rest	Rest	Rest

Table 1.3: Relation between emotional dimensions and categories for the system proposed in [García and Alías, 2008].

1.5.2 Data-driven classification

This section contemplates the classification strategies that are somehow based on the information extracted from evidence that is compiled in a training corpus. This information is usually used to automatically build the classifier of use. Within this classification approach, it is useful to differentiate the deterministic strategies from the probabilistic ones. These aspects are treated next.

Deterministic approach

Deterministic classifiers assume or fit a discriminant function to separate the training examples in the feature space and decide on the result yielded by some similarity measure. These classifiers interpret distances as a sense of membership (when elements are close together) or dissimilarity (the inverse).

Previous work with deterministic classifiers for TSP may include Support Vector Machines (SVM), which apply a maximum-distance criterion among the categories to assess. SVM are especially robust to noisy data because only the support vectors are effective for decision making [Lan et al., 2009]. Although in general SVM yield very good results, when there are not enough examples to represent the training space accurately they decrease their performance dramatically being even unable to operate [Sassano, 2003]. In order to avoid this problem, other approaches may be proposed, such as linear classifiers in feature space like the Winnow update rule [Alm et al., 2005], or through the application of dimensionality reduction techniques like Latent Semantic Analysis [Strapparava and Mihalcea, 2008].

As an illustrative example of deterministic approach to TSP, firstly on the circumplex, this work refers to the *centroid-based* classifier for the interpretable results that it yields. Given the dimensions for a set of examples pertaining to the same category, the centroid (mean) is computed and applied to determine whether an unseen example pertains to the same category through a minum-distance criterion. Its definition in real-valued space is in concordance with the gradual nature of emotional dimensions [Schröder, 2004a]. The centroids are learnt from the average dimensions of training examples according to their gold-standard sentiment labels.

Secondly, on the weighted VSM, it is referred to the *Associative Relational Network* (ARN) [Alías et al., 2008]. The ARN can be regarded as a graph-based model for generic text representation that includes all words

and their order in the text (see Figure 1.4). It works under the *distributional hypothesis*, where words found in similar contexts tend to be semantically similar [Firth, 1957], and therefore, its semantic orientation is also expected to be similar.

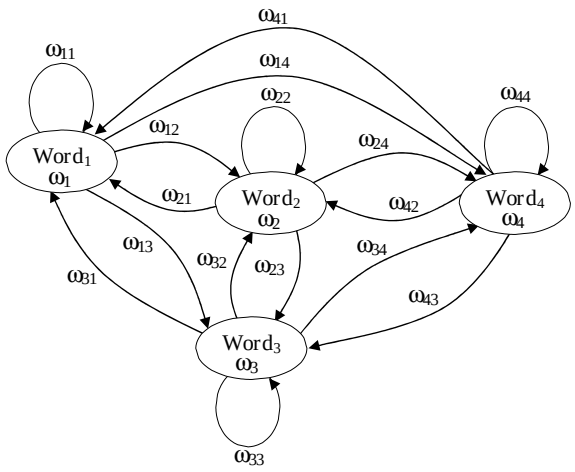

Figure 1.4: The structure and sequentiality of text in the Associative Relational Network [Alías et al., 2008].

The ARN includes some stylistic features by pairwise coupling all observable words in the training corpus, building a graph similar to Figure 1.4. Then, the resulting graph may be vectorised into a VSM (the words and the links of words constituting the dimensions of the VSM). This space may be used for classification purposes in the so called Full ARN (ARN-F). Nonetheless, better results are reported in [Alías et al., 2008] when the feature space used for evaluation is defined only by the words observed in the texts to test (Reduced ARN, or ARN-R), as the classification space sparseness is drastically reduced (the relative performance increase of the ARN-R wrt the ARN-F is around 19% in average). Additionally, the regarded pairing of words in order, aka *tuples*, may be a collocation [Manning and Schütze, 1999], i.e. the appearance of two consecutive words that lose their meaning if separated (e.g., phrasal verbs, certain expressions, etc). If so, their conventional special behaviour, e.g., some noun phrases or phrasal verbs, can only be grasped through the consideration of their sequential existence as a single term, along with the words that make the tuple, a move formerly taken in [Pang et al., 2002].

Probabilistic approach

In [Manning and Schütze, 1999] the authors state that cognitive and linguistic procedures are better explained probabilistically. According to their insight, sentiment recognition systems should better be designed with probabilistic models in mind (trained with statistical evidence). But other authors critique statistical Natural Language Processing for being unsuccessful at sentence-level and introduce common-sense knowledge heuristics extracted from selected databases (see a compilation of some of them in [Alm et al., 2005]).

Previous works with probabilistic classifiers may include n-grams, which model the distribution of word-based n-tuples. Although their nature may be appropriate to build models of style, their statistical robustness may result low due to a shortfall of training data [Alías et al., 2008] (a problem equivalent to the one affecting SVM). Nevertheless, unigrams and bigrams are used with success in a different environment (movie reviews) [Pang et al., 2002], where they are considered to be an orthogonal way to incorporate context.

Now the Naive Bayes (NB) *generative* classifier is reviewed as it best exemplifies the solid philosophical foundation of probability. Furthermore, its flexibility allows NB to be applied on a textual environment [Pang et al., 2002] as well as a real-valued environment [Witten and Frank, 2005]. Therefore, it is suitable to be used on emotional dimensions as well as term-weighted features. Different environments present different properties, and therefore entail using different strategies. Thus, for the former case (textual environment), the models may be built with Bernoulli distributions, while for the latter (real-valued environment) they may be Gaussian densities [Manning and Klein, 2003]. With regard to the independence assumption of the NB, it has been shown that the dimensions of the circumplex are considered to be nearly independent [Mehrabian, 1995], which makes the NB a suitable choice in this case. However, the textual features cannot fit the same assumption since their appearance is somewhat established by a set of grammatical norms.

Finally, the application of the MaxEnt classifier, aka Multivariate Logistic Regression [Manning and Schütze, 1999, Manning and Klein, 2003], is reexamined as a representative *discriminative* probabilistic classifier in TSP [Pang et al., 2002]. MaxEnt models do not assume any conditional relation among the features, and so might potentially perform better when conditional independence assumptions are not met, that is the situation for

textual features as aforementioned. Even with exactly the same features, changing from joint to conditional estimation may increase performance [Manning and Klein, 2003]. Moreover, its purpose of maintaining the model as uniformly distributed as possible, only applying constraints evidenced in the training corpora, makes it an especially fair classifier. In TSP, MaxEnt models are traditionally defined to deal with binary data [Pang et al., 2002], therefore text presence, not with the real-valued data of the circumplex, although a recent proposal [Yu et al., 2009] may surpass this restriction using a training algorithm based on spline interpolation.

1.6 Hierarchical and risk-assessed strategies

The criterion to split the classes in the hierarchy has a direct effect on the structure of the classifier. For example, [Wilson et al., 2009] first separates the neutral utterances and then decides the polarity on the non-neutral ones, whereas [Alm et al., 2005] also contemplates the assessment of all the classes together. This criterion divergence implies that the splitting thresholds (parameters of the classifiers) will be tuned to different values.

What is addressed in this section is the problem division into *decision levels* (hierarchy) and the assessment of the *risk* of incurring an error. On the one hand, it is easier to differentiate items/concepts at the top of the hierarchy as they are most dissimilar while distinctions are more difficult in descendant order because they are more similar (as it happens with the entropy reduction of a decision tree [Duda et al., 2000]). On the other hand, if the decision is taken in one single step, all categories are observable at the time of decision (unlike the hierarchy), and it can be pondered how venturesome is to decide on one category wrt the others. These two facts around the decision process may be reflected in the hierarchical and risk-assessed strategies explained below.

1.6.1 Minimum-risk: posterior weighting strategy

Generally some of the consequences of a wrong decision may be more adverse than others. Given a hierarchical classification of emotion [Shaver et al., 1987], a classification mistake within the same overall sentiment evaluation (take “sadness” for “anger”) may not be as critical as an error among the overall sentiments (take “sadness” for “joy”) in the con-

text of speech synthesis. There exists an inherent bond between emotions and their overall sentiment polarity. In terms of distances in the emotion hierarchy this statement has a point.

In decision-theoretic terminology, the expected loss associated to misclassification is called a *risk* [Duda et al., 2000], and consequently this risk may be set conditional upon a decision. Equation (1.4) defines the conditional risk function. $R(c|\mathbf{x})$ represents the risk associated to choosing class c while observing feature vector $\mathbf{x}$, λ_k^c represents the cost associated to deciding on class k while analysing the risk of selecting class c, and $P(k|\mathbf{x})$ represents the posterior probability yielded by a plain Bayesian decisor (with typical Gaussian likelihood models).

$$R(c|\mathbf{x}) = \sum_k \lambda_k^c \; P(k|\mathbf{x}) \tag{1.4}$$

The resulting minimum overall risk shown in Eq. (1.5) is called the *Bayes risk* and is the best performance that can be achieved with a Bayesian decison strategy. This minimum corresponds to the class c' that is most likely to be representative for the feature vector $\mathbf{x}$. This is the classification criterion.

$$R(c'|\mathbf{x}) = \arg\min_c \left[R(c|\mathbf{x})\right] \tag{1.5}$$

Note in Eq. (1.5) that if all errors are equally costly, that is, the loss incurred for any misclassification is the same, the minimum-risk decision rule yields the same conclusions as the plain Bayesian decisor. Anyhow, in the general case the loss incurred for making an error is greater than the loss incurred for being correct. Thus in practise, the decision is generally determined by the most likely sentiment tag, and the minor details pointed out wrt the emotion hierarchy can be reflected in the weight of the costs (especially interesting in doubtful situations).

In this context, the differences or distances between classes may be adjusted accordingly in the cost domain, thus embodying the hierarchical structure of emotion (distances among the categories in the taxonomy). Additionally, this approach is interesting for TTS synthesis as it is preferable to choose a neutral speaking style in case of doubt instead of making a huge mistake (take 'joy' for 'sadness').

As a means of weight adjustment, given the classification problem at hand, there are as many optimal parametrisations as number of risk functions, that is one for each class. The target risk, i.e., the desired output of

the risk function, according to a given class c is defined as $t_c(\mathbf{x})$. This target should be set according to a weighted criterion wrt the emotion hierarchy (along with the TTS synthesis criterion).

The global cost function defined in Eq. (1.6), which corresponds to the sum-of-squared errors over all classes, is optimised through the minimisation of the class-dependent cost functions (see Eq. (1.7)).

$$J(\lambda) = \sum_{c} \left[J_c(\lambda^c) \right] \tag{1.6}$$

$$J_c(\lambda^c) = \sum_{\mathbf{x}} \left[R(c|\mathbf{x}; \lambda^c) - t_c(\mathbf{x}) \right]^2 \tag{1.7}$$

The Minimum Squared Error procedure establishes that the squared-error function $J_c(\lambda^c)$ under study is to be minimised. In the particular case of Eq. (1.7) the error surface may easily have a multiplicity of minima due to the weighted sum of Gaussians in the risk function. As a solution it is proposed a gradient descent procedure where the update rule accounts for this situation and incorporates a momentum parameter α, see Eq. (1.8). The weight increment provided by the momentum should bypass sharp local minima and small plateaus but stay at the absolute minimum.

$$\begin{aligned} \lambda^c(m+1) = \; & \lambda^c(m) \\ & - (1-\alpha)\eta(m)\nabla J_c(\lambda^c(m)) \\ & + \alpha(\lambda^c(m) - \lambda^c(m-1)) \end{aligned} \tag{1.8}$$

The contribution of the gradient in Eq. (1.8) for a given component k can be resolved into Eq. (1.9).

$$\frac{\partial}{\partial \lambda_k^c} J_c(\lambda^c) = \sum_{\mathbf{x}} 2 \left[R(c|\mathbf{x}; \lambda^c) - t_c(\mathbf{x}) \right] P(k|\mathbf{x}) \tag{1.9}$$

Note in Eq. (1.9) that each weight update requires the computation over all training samples. This way, the trajectory of the weight vector is smoothed, but it may take the algorithm a long while to converge. As an alternative, a single-sample error correction procedure is proposed: one training example is computed at a time. Eventually $\lambda^c(m)$ should converge to a limit vector on the boundary of the solution region, remaining there for all m greater than some finite value. This bound is specified with a threshold and highly influenced by the learning rate $\eta(m)$, which determines the step size.

As it can be seen, the problem at hand is not linearly separable, like the majority of Natural Language Processing problems [Manning and Schütze, 1999]. In these cases, the use a step-variable learning rate is advised to ensure error convergence. Therefore, $\eta(m) = \frac{1}{m}$ is assured to provide a good learning behaviour [Duda et al., 2000]. Notice that the convergence in time steps of the learning rate implies that the contribution of the last instances is less important than the first ones. Since some examples may be more representative than others, this work proposes to iterate over random training examples until the weights converge. Following this proposal the risk of obtaining a bad solution by accidentally accepting the result at an unfortunate termination time may be reduced.

Nevertheless, all this gradient descent discussion is tightly subject to the data at hand, and for a variety of reasons (but mainly due to a shortage of data) this nice implementation might not work. In those cases, a solution would consider the linear difference (i.e., the distances) among the affective classes. Since this risk-assessment approach is based on posterior probabilities, the distances defined in another space, say the circumplex, are a good choice to scale the posteriors: longer distances imply more dissimilarity, and therefore a higher risk.

1.6.2 Decision levels: hierarchical strategy

The hierarchical approach follows the natural *taxonomy* of emotion and assesses the different levels of cognitive difficulty in sentiment categorisation [Shaver et al., 1987]. This graded difficulty approach has a likelihood with the entropy gains used in decision trees to decide on the thresholds that define the structure of the trees. While this is the theory-compliant approach, other practical strategies, e.g., the Optimal Stacks of Binary Classifiers, set a pairwise coupling with all the classes [Lan et al., 2009, Koppel and Schler, 2006]. This method treats the constituent pairwise problems identically and may obtain good results although sometimes produces counter-intuitive uneven rules (bound to the nature of the dataset at hand). Thus, the taxonomy emotion [Shaver et al., 1987] integrated in the polarity hierarchy used in [Murray and Carenini, 2009], see Figure 1.5, seems to best capture the notion of hierarchical dependence between emotions and sentiments.

One of the main benefits of addressing the "neutral" class in the first level of the hierarchy responds to the fact that the best way to improve performance over all polarity classes, i.e., "positive" and "negative", is to improve

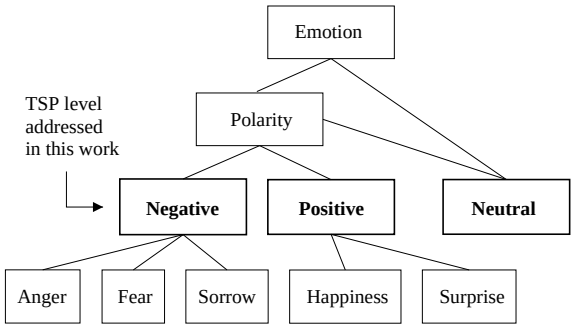

Figure 1.5: Taxonomy of emotion considered in this work. Although the main purpose is to attain sentiment classification, emotions are also shown here for conceptual clarity.

the system's ability to identify neutral instances [Wilson et al., 2009]. Finally note the double transition to the "neutral" sentiment. If it comes from the root of the taxonomy then the text in question does not contain any emotionated word, but if comes from the "polarity", then the text does actually contain affective words but the sentiment is neutrally balanced on average.

1.7 Experiments

The assessment of the whole diversity of methods to tackle the tripartite task of sentiment classification from text (i.e., positive/negative/neutral) defines a framework full of possible configurations. In order to give an overview of the applicability of the reviewed strategies to TSP, some that exemplify the philosophy behind each proposal are submitted to experimentation on two different datasets. The experiments intend to deal with different aspects of the TSP problem, treating the relevant features (based on their textual form or represented in a circumplex) and the classification principles (probabilistic or deterministic).

1.7.1 Sentiment datasets

This section describes the selected datasets and prepares them for the sentiment classification task. If the dataset in question is not labelled with the classes covered in this dissertation (i.e., positive/negative/neutral), then some strategy is follow to adequate it to the interests pursued in this work.

Headline	A	D	F	J	Sa	Su
Bombers kill shoppers.	66	39	94	0	86	0
Man rides stationary bike for 85 hours.	0	0	0	18	0	87

Table 1.4: Tagging examples of the Semeval 2007 dataset. "A" stands for Anger, "D" for Disgust, "F" for Fear, "J" for Joy, "Sa" for Sadness and "Su" for Surprise.

Semeval 2007 dataset

The first corpus of use is the whole Semeval 2007 dataset [Strapparava and Mihalcea, 2007]. It consists of a compilation of 1250 news headlines, as the authors state that this kind of data is produced to arise feelings in the readers.

These headlines were appraised in six different emotions by different evaluators, as reported by the emotion labelling task described in [Strapparava and Mihalcea, 2007] (also see [Strapparava and Mihalcea, 2008]). The six emotions considered were weighted according to their individual contribution to each headline. For example, see Table 1.4. Since this work considers sentiment categories, there is a conversion process required to treat the dataset accordingly. As a first step, this issue was tackled by somewhat heuristic decisions experimentally validated [García and Alías, 2008]. However, there is room for further improvements.

As a next step, this dissertation proposes considering formal emotional theories based on the circumplex to map the Semeval 2007 headlines and obtain their representation in a space of emotional dimensions. Since there is no unified theory of affect, three different emotion representations are considered to find the best one for mapping the Semeval 2007 dataset: Russell's model, Whissell's dictionary and Scherer's theory of affect. Note that not all the basic emotions in the Semeval 2007 dataset can be directly mapped into the emotional representations proposed (for example, "disgust", "joy" and "surprise" cannot be found in Russell's model). In order to surpass this mismatch, the existing similarity between two emotions close together in the circumplex model is used [Schröder, 2004b], accounting for the synonyms for each emotion given by WordNet [Fellbaum, 1998] (with respect to the pre-

vious example, “annoyed” is taken for “disgusted”, “delighted” is taken for “joyful” and “astonished” is taken for “surprised”). The resulting distribution of basic emotions in the circumplex according to Russell’s model is shown in Figure 1.6, and according to Whissell’s dictionary in Figure 1.7. These two example model distributions in the circumplex are shown to display their mismatch and the expected difference in mapping the dataset.

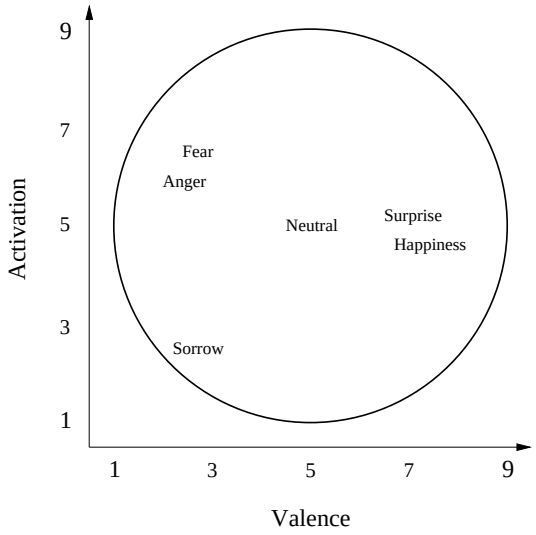

Figure 1.6: Distribution of basic emotion categories in the circumplex according to Russell’s model of affect.

Due to the availability of the dataset, both the training and test sets (1250 headlines in total) are considered for conducting the sentiment analysis experiments. Taking each annotation of emotion (out of the six annotations for each headline) for a weighed vector, the vector sum can be computed in order to obtain the resulting projection of the headline in the given emotional space (circumplex). A similar approach was followed in [Hofer et al., 2005]. Then the closest basic emotion to this resulting point is assigned to the headline.

In order to score the adequacy of the affective model to map the dataset into a space of emotional dimensions, a 10-fold cross-validation procedure with an example-based classifier should yield the effectiveness measure of the dataset wrt a model of affect. A classifier based solely on previous examples should miss the advantages of a model produced inductively, and thus it should not be biased by the different strategies available for this training

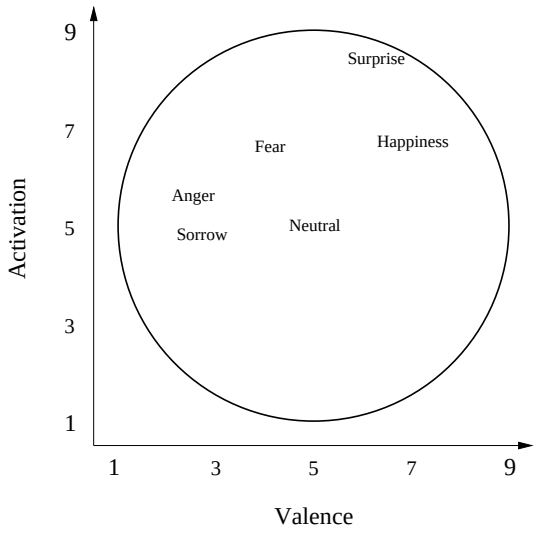

Figure 1.7: Distribution of basic emotion categories in the circumplex according to Whissell's dictionary of affect.

purpose. In this sense, a k-Nearest Neighbour is an adequate classifier candidate. Regarding the number of neighbours considered (i.e., the k), which determines the smoothness of the boundaries as k increases (also preventing overtraining) it is set to 7 as this number of neighbours yields quite smooth boundaries. Therefore, a 7-Nearest Neighbour (7-NN) should be a general enough classifier to decide the best model of affect for mapping the dataset. By following this procedure, it should be seen with which model of affect the similar headlines are best grouped. After dealing with the three proposed models, the one which results in a highest effectiveness rate for mapping the dataset at hand, and thus building the ground truth, will be taken for further analysis. Notice that the classification performance is computed by means of the macroaveraged F_1^M measure [Sebastiani and Ricerche, 2002] so as to prevent the results from being biased due to the balance of the data distribution.

Labelling each headline in the dataset with the nearest emotion given by a determined model of affect, Table 1.5 can be produced showing the resulting balance of the Semeval 2007 dataset regarding each model. For Whissell's model and Scherer's model some emotions are barely represented, while for Russell's model all emotions have a reasonable amount of instances.

As it can be seen in Table 1.5 the representation of the Semeval 2007

Table 1.5: Distribution of emotions in the Semeval 2007 dataset according to the considered models of affect.

Emotion	Russell	Whissell	Scherer
anger	21.55 %	14.02 %	12.58 %
fear	6.09 %	37.42 %	8.89 %
sorrow	5.69 %	0.32 %	1.92 %
neutral	53.93 %	20.03 %	54.41 %
happiness	9.21 %	25.64 %	15.30 %
surprise	3.53 %	2.56 %	6.89 %

dataset in the different emotional spaces shows unbalanced distributions. In the 10-fold cross-validation procedure, see Table 1.6, for Whissell's dictionary of affect the 7-NN classifier is unable to predict the categories (emotions) with the lowest generality [Sebastiani and Ricerche, 2002] (i.e. scarcely populated) due to the lack of examples. On the contrary, Russell's model and Scherer's perform successfully. These two models are very alike (their difference is not statistically significant).

Table 1.6: Adequacy of the considered models of affect for mapping the Semeval 2007 dataset.

Model of affect	Macroaveraged F_1 (mean $\pm$ std)
Russell	96.78% $\pm$ 2.52
Whissell	N/A
Scherer	95.35% $\pm$ 2.66

By little difference, Russell's model of affect has resulted to be the best affective model to represent the emotions of the dataset at hand considering each emotion separately. Thus, it is chosen to label the emotions of the Semeval 2007 dataset. Nevertheless, the aimed task in this dissertation is sentiment classification. Therefore, the final sentiments tags associated to the headlines of this dataset are yielded by grouping all the "anger", "fear" and "sorrow" instances into a "negative" class and all the "happiness" and "surprise" instances into the complementary "positive" class, leaving the neutral instances with the "neutral" sentiment tag.

Property	Dataset	
	Semeval 2007	FWF corpus
Vocabulary	3773	2771
4-lexicon	405	304
Total lexicon	10087	8523
Perplexity	924.13	624.82
Instances	1250	758
# neg.	33.33%	22.67%
# pos.	12.20%	10.67%
# neu.	54.47%	66.66%

Table 1.7: Properties of the sentiment datasets. 4-lexicon represents the lexicon of words appearing at least 4 times.

Fifty Word Fiction (FWF) corpus

The second dataset is the Fifty Word Fiction (FWF) corpus [Read, 2004] which is a collection of 155 fifty-word long stories, with 758 sentences in all. The author points out that the brevity of the stories potentially compelled the writers to use highly affective language. Each sentence was manually annotated by several human evaluators for its sentiment ("positive", "negative" or "unclassifiable") and according to a psychological model of affect [Watson and Tellegen, 1985]. In order to be consistent with the discourse on the sentiment labels used in this dissertation, an instance which is unclear wrt its polarity, i.e., "unclassifiable" in the FWF corpus, is considered to pertain to the "neutral" category. It must be stated that "unclassifiable" and "neutral" are not the same concepts, but given that the other two sentiments are "positive" and "negative", "neutral" seems to be the most likely tag for "unclassifiable" for the tripartite classification task at hand.

Corpora characteristics

Table 1.7 shows some properties of the datasets.

Note in passing that the datasets pertain to different *domains*, as one topic is news headlines while the other is fiction stories. What is more, the two models of affect used for the gold-standard labelling process also dissent. In sum, all these differences will make it difficult to generalise the conclusions,

a problem commonly identified as the *sentiment transfer*, and dealing with it deserves special attention [Li and Zong, 2008]. In any case, the results obtained with each dataset will serve as an overall performance in TSP.

1.7.2 TSP schemes

Several TSP systems are considered for sentiment analysis, given the diversity of approaches at hand. For the ones that make use of the circumplex, the ANEW dictionary of affect is used because it is general-purpose and it was created specifically by a psychological study measuring the associations between words and human emotions. And for the approaches that make use of raw text directly, it is considered using a VSM weighted with TW metrics based on frequencies to represent the textual features.

Regarding the principles of classification, firstly the focus is on the heuristic-rules approach given the previous work on EmoLib. In contrast, its performance is compared to some data-driven approaches. With regard to the deterministic approach, the Nearest Centroid (NC) classifier is proposed for the circumplex because of its generality [Lan et al., 2009] and ease of interpretation. Next, for the weighted VSM the application of ARN-R is suggested. This scheme has already exhibited its power to be more effective at the one-sentence level text classification job than other reasonable techniques like k-Nearest Neighbours (example-based), Independent Component Analysis based (predominantly thematic approach) and bigrams (at character-level, due to a lack of enough training examples at word-level) [Alías et al., 2008]. Lastly, the probabilistic approach. With respect to the circumplex it is proposed NB with Gaussian density likelihood distributions, for its consistence with certain axioms of rational inference. And in regard of the textual feature approach, on the one hand, NB with Bernoulli distributions is evaluated, and on the other hand, MaxEnt is considered as dealing with conditional probabilities may increase the classification performance [Manning and Klein, 2003].

Next, the hierarchical and risk-assessed strategies are evaluated. It must be considered, though, that the application of these strategies is bound to the nature of the feature space of use. Therefore, the hierarchical strategy is applied on the VSM and the risk-assessed strategy on the circumplex. Otherwise, the strategies make little sense in some cases. For example, the hierarchy would collapse the polar instances to a location very close to "neutral" in the circumplex, as "neutral" lies at an intermediate position between

"negative" and "positive" in the emotional dimensional space. In the VSM, this fact is less critical as the principal components for each sentiment category could be computed and used to assess some risk. Nevertheless, this last proposal is not evaluated in this dissertation and is left for future work. And for practical purposes, preliminary results with the gradient descent procedure to adjust the risk costs have yielded poor results as the values they have converged have not been much different from their random initialisation. Thus, the distances with the rest of the sentiment centroids are taken for risk weights.

Finally, one remark on SVM and n-grams, two of the most promising schemes to TSP and their poor applicability to the considered datasets. In [Joachims, 2006] it is reported that an ordinal regression[3] SVM over n examples (size of the test set) is solved by translating it into a classification SVM with $O(n^2)$ examples. This requirement is impossible to meet when the needed number of examples exceeds the size of the training corpus by an order of magnitude, see Table 1.7. And for bigrams, for instance, given a vocabulary with the words that occur at least 4 times, the average rate of observed bigrams is 7.69%, covering 10.85% of the total vocabulary, which indicates that this model is impracticable to the datasets.

1.8 Results and discussion

To gauge the efficacy of the various classification schemes (features + evaluation space + classifier) this work relies on the macroaveraged F_1^M measure [Sebastiani and Ricerche, 2002], and in order to estimate their performed effectiveness, the 10-fold cross-validation procedure is applied to each experiment at sentence-level for the data-driven approaches (highlighting its mean and std in the results). For the heuristic approach evaluated, since no data training is needed, the F_1^M measure computed on the whole corpus is directly shown (no std). Where possible, the comparisons are evaluated for statistical significance at the 0.05 confidence level through ANOVA tests.

[3]Sense of grading on scales likely to be present in IR issues.

Exp. num.	**Features**	**Classifier**	F_1^M **measure (mean ± std)**	
			Semeval 2007	FWF corpus
1	3D emotional dimensions	Heuristic rules	48.34%	N/A
2		Nearest Centroid	48.85% ± 4.23	**45.05% ± 5.17**
3		Gaussian NB (GNB)	46.00% ± 4.77	44.09% ± 4.54
4		Risk-weighted GNB	**50.31% ± 5.44**	40.83% ± 8.69
5	Term freq.	ARN-R	46.68% ± 4.10	35.96% ± 5.95
6	Term freq. + tuple freq.	ARN-R	47.81% ± 3.74	33.48% ± 7.31
7	Inverse term freq.	ARN-R	41.58% ± 4.73	36.74% ± 5.29
8	Inverse term freq. + tuple freq.	ARN-R	45.93% ± 6.02	37.25% ± 5.43
9	Relevance factor	ARN-R	56.95% ± 7.95	N/A
10	Relevance factor + tuple freq.	ARN-R	58.09% ± 7.17	N/A
11	Binary term pres.	Bernoulli Naive Bayes	12.69% ± 2.37	15.84% ± 7.58
12	Binary term pres. + tuple pres.	Bernoulli Naive Bayes	13.70% ± 0.97	24.33% (*)
13	Binary term pres.	Maximum Entropy	51.94% ± 8.41	**40.39% ± 4.57**
14	Binary term pres. + tuple pres.	Maximum Entropy	52.59% ± 7.40	39.77% ± 4.95
15	Term freq.	Hierarchical ARN-R	46.01% ± 3.34	36.18% ± 6.08
16	Term freq. + tuple freq.	Hierarchical ARN-R	47.60% ± 3.88	35.11% ± 6.85
17	Inverse term freq.	Hierarchical ARN-R	43.70% ± 4.53	34.09% ± 6.27
18	Inverse term freq. + tuple freq.	Hierarchical ARN-R	47.46% ± 7.54	36.42% ± 6.02
19	Relevance factor	Hierarchical ARN-R	53.04% ± 8.70	N/A
20	Relevance factor + tuple freq.	Hierarchical ARN-R	**58.20% ± 6.13**	23.77% (*)

(*) Result yielded by one single fold. Effectiveness rates (precision in particular) for the other nine folds were incomputable.

Table 1.8: Results from the comparative study. Terms are considered to be single words. The two best results for each corpus are printed in boldface, one for emotional dimensions and the other for textual features.

1.8.1 Dataset comparison

With respect to the results from the comparative study, see Table 1.8, at first sight the overall F_1^M results obtained for the Semeval 2007 are somewhat better than for the FWF corpus (10.49% better on average for the statistically significant experiments 4–8 and 13–18). A straight explanation to this would be the data shortage in FWF. Fewer examples of the sentiment category with the least generality (roughly the half), that is the "positive", sharply penalise the effectiveness of the classification strategy. However, this comparison is too abstract because many different factors are involved in the considered results, and thus it can hardly be told whether this corpora difference is due to the properties of the datasets or to the characteristics of the classification strategies.

Note the perplexity difference between the datasets, being the Semeval 2007 dataset a more entropic (i.e., the log inverse of the perplexity) dataset than the FWF corpus, see Table 1.7. The greater the perplexity in the corpora, the greater the amount of information it contains. This trait should be regarded as positive for classification purposes, recall the maximum entropy classification criterion.

1.8.2 Features comparison

Note that most experiments (ten out of eleven) that deal with textual data (i.e., dealing with the topic-dependent words of each corpus domain) show a statistically significant between the datasets (experiments 5–8 and 13–18), whereas only one experiment (out of three) dealing with emotional dimensions (experiment 4), shows a statistically significant difference between the corpora (yielding an effectiveness rate comparable to the best experiments with textual data). This fact may expose the data-domain (and maybe data-size) independence of a general-purpose dictionary of affect with emotional dimensions, and in contrast, the great dependence of the approaches that deal with textual features, despite scoring the best effectiveness rates.

Regarding the emotional dimensions it is observed that the independent normal distribution of the emotional dimensions is an acceptable assumption wrt simple mean distance since no statistically significant difference is observed between experiments 2 and 3. That is, the fact of considering the variance of data in addition to its mean, and assuming that the data is distributed normally, has an almost insignificant effect on the results, despite

they seem to be slightly worse.

On the textual feature approaches (experiments 5–20), the consideration of tuple frequencies wrt single term frequencies might seem to meliorate the effectiveness rate, but the improvements are not statistically significant. As the observation of tuples may be more singular wrt a sentiment category than words alone, its weighting should be raised with the RF. Nevertheless, the good results obtained with RF for one dataset (Semeval 2007) are greatly decreased for the other (FWF), see experiments 9, 10 and 19, being unable to be computed in some cases. This could again be attributed to a shortfall of frequent words and/or their distribution among the sentiment classes, e.g., see Table 1.9 and Table 1.10. Note that *all* the words are considered, including punctuation marks, and splitting on them. This responds to the aim of grasping the stylistic properties of text [Alías et al., 2008]. According to these tables, the distribution of frequent words alone is of little help to discern the sentiments (most of them are biased towards the neutral sentiment on the two datasets, that is the category with the greatest number of examples), whereas the distribution of frequent tuples presents different results. In particular, they are biased towards the negative sentiment for the Semeval 2007 dataset, while the bias towards the neutral remains for the FWF corpus. These facts may help understand why the effectiveness rates are generally lower for the FWF corpus. Nevertheless, a more thorough study would be necessary to extract further conclusions.

Observe as well that the use of ITF wrt TF performs slightly bad (5.09% decrease with the statistically significant experiment 7 wrt experiment 5 for Semeval 2007), contrary to the results in traditional TC reported in [Alías et al., 2008], although a corpus built with topic-dependent sentences was used. Perhaps this dependency with the topic is crucial for the ITF weighting factor.

1.8.3 Classifier comparison

Firstly, relating to the previous work on EmoLib, i.e. the set of heuristic rules produced wrt the Semeval 2007 dataset [García and Alías, 2008], it can be seen how this approach (experiment 1) yields a baseline better than the random choice for three classes. Nevertheless, since this system was heuristically produced wrt the Semeval 2007 dataset, it cannot be tested on the FWF corpus unless some new set of rules wrt the FWF corpus are expertly defined. This job is though outside the scope of this research work.

Word/tuple w	**Corpus counts**	$P(\mathbf{N}\|w)$	$P(\mathbf{neutral}\|w)$	$P(\mathbf{P}\|w)$
.	1205	33.69%	54.02%	12.28%
’	216	23.61%	62.03%	14.35%
in	197	48.22%	44.16%	7.61%
to	171	33.33%	59.64%	7.01%
,	162	38.88%	50.61%	10.49%
:	120	24.16%	61.66%	14.16%
s	116	19.82%	60.34%	19.82%
for	114	27.19%	60.52%	12.28%
-	105	33.33%	59.04%	7.61%
on	99	34.34%	53.53%	12.12%
’ s	116	19.82%	60.34%	19.82%
S .	35	54.28%	40.00%	5.71%
U .	35	68.18%	22.72%	9.09%
’ .	34	20.98%	62.96%	16.04%
. S	34	55.88%	38.23%	5.88%
North Korea	18	41.66%	50.00%	8.33%
in Iraq	14	85.71%	14.28%	0%
Iraq .	13	61.11%	38.88%	0%
’ t	12	25.00%	66.66%	8.33%
says .	11	45.45%	36.36%	18.18%

Table 1.9: List of the ten most frequent words and tuples and their distribution in the Semeval 2007 dataset, expressed in terms of conditional probability percentage (approximated).

Word/tuple w	**Corpus counts**	$P(\mathbf{N}\|w)$	$P(\mathbf{neutral}\|w)$	$P(\mathbf{P}\|w)$
.	665	22.85%	67.21%	9.92%
,	307	28.33%	58.30%	13.35%
the	279	23.29%	64.51%	12.18%
and	167	23.95%	62.87%	13.17%
a	156	22.43%	57.69%	19.87%
”	145	25.00%	61.80%	13.19%
I	136	34.93%	51.80%	13.25%
to	117	26.49%	59.82%	13.67%
of	110	22.72%	59.09%	18.18%
in	92	20.65%	64.13%	15.21%
. ”	66	25.39%	61.90%	12.69%
? ”	35	8.57%	77.14%	14.28%
in the	27	11.11%	74.07%	14.81%
, ”	26	30.76%	46.15%	23.07%
on the	24	33.33%	41.66%	25.00%
, and	21	23.80%	47.61%	28.57%
, the	20	25.00%	65.00%	10.00%
of the	17	5.88%	76.47%	17.64%
, he	16	37.50%	56.25%	6.25%
at the	16	6.66%	80.00%	13.33%

Table 1.10: List of the ten most frequent words and tuples and their distribution in the FWF corpus, expressed in terms of conditional probability percentage (approximated).

In contrast, the data-driven strategies can be automatically adapted to new environments (datasets). For deterministic approaches (experiments 2, 5–10 and 15–20) no statistically significant difference between results is observed, whereas for probabilistic approaches (experiments 3, 4 and 11–14) a large statistically significant difference is observed between the Bernoulli Naive Bayes and the Maximum Entropy[4] approaches. In the context of these experiments, keeping the model as uniform as possible (the maximum entropy principle) is a smart move for classification purposes (34.23% of improvement).

Next, the purpose is to analyse if probabilistic models take advantage from class-likelihood distributions to assess data that was unseen at the training stage, and thus may perform more gracefully than deterministic models in a general context (note that the hierarchical strategies with the ARN-R are also considered despite the heuristic definition of the hierarchy). The obtained results show no statistically significant difference between the Gaussian models and the centroid models (experiment 3 wrt experiment 2), nor the difference between the best results with MaxEnt models and the ARN models (experiment 14 wrt experiment 10). Even in the latter case the deterministic approach (ARN-R) seems to have performed somewhat better than the probabilistic approach (MaxEnt). So the theoretic advantage of building class-likelihood distributions in front of unseen features is unclear for the data considered in this work.

While comparisons among the principles of classification are difficult because of the different information that the features provide, the differences shown by the best systems in Table 1.8 (in boldface) are statistically significant between experiments 4 and 20 for Semeval 2007 and between experiments 2 and 13 for FWF. In the first case a Hierarchical ARN-R (deterministic classifier) performs significantly better than a Risk-weighted NB (probabilistic classifier) with a 7.89% of improvement, and in the second case a Nearest Centroid (deterministic) works significantly better than a MaxEnt classifier (probabilistic) with a 4.66% of improvement. In both cases, the deterministic principle has shown better performance than the probabilistic principle. For one dataset (the one with a larger list of frequent words, that is the Semeval 2007 dataset) it works better with textual features while for the other (that is the FWF corpus) it works better with emotional dimensions. Further experiments with more corpora would be desirable to assert

[4]http://nlp.stanford.edu/downloads/classifier.shtml

this statement, though. Otherwise it's quite a venturesome argument that only applies to the datasets considered.

1.8.4 Strategy comparison

Last, the incorporation of the risk-assessment or the hierarchy (structural refinements) has led to interesting results, although they are not shared for the two datasets. In particular, all the results are worse for FWF than for Semeval 2007. For example, the best results for Semeval 2007 (one with emotional dimensions and the other with textual features) almost correspond to the worst results for FWF and the difference is statistically significant in experiment 4). This fact leads to the tentative belief that the successful incorporation of these strategies is dependent on the size of the corpus: the larger the dataset, the better.

Nevertheless, none of the performance improvements is statistically significant wrt the classification strategy without the structural melioration intent (despite its seems to work somewhat better). Therefore, with the datasets at hand, further conclusions are just probationary. Further research is welcome to assert this discussion.

Chapter 2

Conclusions and future work

One important requirement spotted in the sentiment analysis task is the general need of a big amount of data (for Machine Learning purposes), especially for the approaches that make use of Support Vector Machines (SVM). Given that the purpose of this research is the production of expressive synthetic speech sentences, the texts to treat are motivated to be recorded given an expressive style [Iriondo et al., 2009]. And this process is costly. Thus, the amount of recorded speech is, in general, desired to be the smaller the better. But the smaller the dataset the more difficult it becomes for techniques like the SVM to be applied with success.

Therefore, with the small dataset-size restriction, and according to the research carried out so far in this dissertation, the most effective directions to Text Sentiment Prediction on the tripartite sentiment recognition task are the Hierarchical Associative Relational Network – Reduced (ARN-R) for textual features and the Risk-weighted Gaussian Naive Bayes for emotional dimensions, as long as the available data is minimally abundant (these conclusions are extracted with the results yielded with the most abundant dataset, i.e. the Semeval 2007, with 1250 sentences). Otherwise, for a greater shortage of data, a MaxEnt classifier for textual features and a Nearest Centroid for emotional dimensions perform better (these conclusions are extracted with the results yielded with the modest dataset, i.e. the Fifty Word Fiction (FWF) corpus, with 758 sentences). The threshold size of the dataset to resolve this compromise remains as an open research question.

Since ARN-based methods yield quite uneven results according to the applied Term Weighting (TW) scheme, especially for the Relevance Factor wrt all the rest, it is sensible to believe that there may be more room for

improvement. By now, this supervised technique has resulted to be the best weighting scheme, but maybe considering other Information Retrieval weights could help it to work better. Moreover, reviewing the basics of ARN-based methods it is found that this technique is "naively" exploited with the evaluation in the VSM, where part of the sequential information is missed in a vector of weighted-terms without a sense of order among its dimensions (a bag-of-words). This aspect may be improved with the consideration of network similarity measures beyond the pattern length introduced in [Alías et al., 2008], which accounts for the number of consecutive terms appearing in the same order in the input text and a given domain. The consideration of all consecutive words in the evaluation of the sentiment is beneficial for the production of synthetic speech by concatenation (in general, the more consecutive words the better in favour of spoken naturalness). Therefore, these kind of network measures are of greatest interest for further research related to speech synthesis. Furthermore, it is also found that a sense of semantic factor is in most need, because the present definition of the ARN cannot deal with negations, for example, whereas other representations like the circumplex enable the pivoting around the location of the "neutral" state.

In the future work it is expected to deal with enhanced features (including the structure) for the ARN-R, that is the best sentiment analysis strategy found in this work that deals with reduced datasets. Other approaches are considered to be explored dealing with larger datasets or through the use of semi-supervised learning techniques [Chapelle et al., 2006], thus being able to use SVM and n-grams. The latter model, apart from language modelling and text classification, has also been used to model subjectivity and other word-based patterns with various levels of lexical instantiation for polarity detection in spontaneous speech [Murray and Carenini, 2009], which would be interesting to investigate. It is also wanted to further explore the capabilities of MaxEnt models and their integration with continuous dimensions and TW schemes [Wang and Acero, 2007].

Likewise, it is desired the inclusion of a Named Entity Recogniser [Nadeau and Sekine, 2007] and a string similarity measure [Islam and Inkpen, 2007] in favour of robustness against the identification of named entities and misspellings when applied to real data, respectively. A preliminary experiment in this context has used the GNU

Aspell[1] spell checker to spot these "word anomalies" and has observed that 2.59% of the words appearing in the corpora (Semeval 2007 and FWF) correspond to these types of word examples. This rate may yield room for future improvement in this direction.

Finally, in Text-to-Speech (TTS) synthesis, when a reader begins reading a story (emphasising the presence of affective data in the text), is makes little sense to account for the parts of the texts that are still to be read. Hence, the consideration of dynamics (sequence model) is more reasonable than the direct static text analysis presented in this work. Emotions have a transitional nature, they blend and overlap along the temporal dimension [Alm et al., 2005]. Accounting for the fact that normally texts are presented in the form of paragraphs (texts longer than a single sentence), in TTS synthesis it would be interesting to assess this temporal evolution of sentiments as a value add of the expressive content conveyed in the spoken voice. This aspect is considered in the future work.

[1] http://aspell.net/

Chapter 3

Contributions

3.1 Scientific publications

Author Alexandre Trilla and Francesc Alías

Title Sentiment classification in English from sentence-level annotations of emotions regarding models of affect

Booktitle Proc. of Interspeech'09

ISSN 1990-9772

Pages 516–519

Address Brighton, UK

Month September

Year 2009

Abstract This paper presents a text classifier for automatically tagging the sentiment of input text according to the emotion that is being conveyed. This system has a pipelined framework composed of Natural Language Processing modules for feature extraction and a hard binary classifier for decision making between positive and negative categories. To do so, the Semeval 2007 dataset composed of sentences emotionally annotated is used for training purposes after being mapped into a model of affect. The resulting scheme stands a first step towards a complete emotion classifier for a future automatic expressive text-to-speech synthesiser.

Note The paper is included in Appendix **??**.

Author Santiago Planet, Ignasi Iriondo, Alexandre Trilla and Francesc Alías

Title Children's Spontaneous Emotion Recognition by Fusion of Acoustic

and Linguistic Features

Booktitle Speech Communication *(In Submission)*

Abstract This paper describes different approaches to perform spontaneous emotion recognition from speech in children. Using the corpus provided for the Interspeech 2009 Emotion Challenge, we propose different classification strategies based on acoustic and linguistic features. Also, we analyse classification structures based on a two-dimensional circumplex model for describing emotions versus structures with no emotional theory basis. Next, we use stacking generalisation to combine the results of these classifiers. Experiments are carried out using leave-one-speaker-out strategy to consider speaker independence. Results show that classifiers non-based on emotional models perform better than those based on the circumplex model, for the corpus at hand. However, their combination outperforms their individual results. Moreover, despite the linguistic classifier results are not good enough to carry out the task of emotion recognition on its own, results are improved when combined with the rest of the classifiers.

Note The text sentiment prediction techniques investigated in this dissertation have been applied in the environment of spontaneous children's speech in order to attain the emotional identification from text, and a contribution has been produced in this submitted publication.

3.2 Associated research projects

This dissertation is partially framed in the context of a funded project that is being developed at the Department of Media Technologies – GTM at La Salle – Universitat Ramon Llull. This project is multidisciplinary as it profits from the knowledge from many scientific fields oriented towards a research line in interfaces for people with disabilities. The description of this projects is shown below:

INREDIS: Interfaces for relation between environment and people with disabilities

- Description: The main objective of this project consists of developing grounding technologies to allow creating communication and interaction channels between disabled people and their environment.

- Name of the funding entity: Ministerio de Industria, Turismo y Comercio (MITyC)
- Origin from the entity: Spanish Government
- Type of entity: Public
- Reference: CEN-2007-2011
- Year funding beginning: 2008
- Year funding end: 2009

Bibliography

[Abbasi et al., 2008] Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Trans. Inf. Syst.*, 26(3):1–34.

[Alías et al., 2008] Alías, F., Sevillano, X., Socoró, J. C., and Gonzalvo, X. (2008). Towards High-Quality Next-Generation Text-to-Speech Synthesis: A Multidomain Approach by Automatic Domain Classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1340–1354.

[Alm et al., 2005] Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proc. of HLT'05*, pages 579–586, Morristown, NJ, USA. ACL.

[Baggia et al., 2008] Baggia, P., Burkhardt, F., Martin, J.-C., Pelachaud, C., Peter, C., Schuller, B., Wilson, I., and Zovato, E. (2008). Elements of an EmotionML 1.0. Technical report, W3C.

[Batliner et al., 2009] Batliner, A., Seppi, D., Steidl, S., and Schuller, B. (2009). Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. *Advances in Human Computer Interaction (AHCI)*.

[Bradley and Lang, 1999] Bradley, M. M. and Lang, P. J. (1999). Affective Norms for English Words (ANEW): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, USA.

[Chapelle et al., 2006] Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA, USA.

[Cowie et al., 2001] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Proc. Mag.*, 18(1):32–80.

[Cruz et al., 2009] Cruz, F., Troyano, J. A., Ortega, J., and Vallejo, C. G. (2009). Inducción de un Lexicón de Opinión Orientado al Dominio. In *Procesamiento del Lenguaje Natural*, number 43, pages 5–12 *(in Spanish)*.

[Duda et al., 2000] Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*. Wiley-Interscience, New York, NY, USA.

[Eide et al., 2004] Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., and Pitrelli, J. (2004). A Corpus-Based Approach to <ahem/> Expressive Speech Synthesis. In *5th ISCA Workshop on Speech Synthesis*, pages 79–84.

[Fellbaum, 1998] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, first edition.

[Firth, 1957] Firth, J. R. (1957). A synopsis of linguistic theory, 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, pages 1–32.

[Francisco and Gervás, 2006] Francisco, V. and Gervás, P. (2006). Exploring the Compositionality of Emotions in Text: Word Emotions, Sentence Emotions and Automated Tagging. In *Proc. of AAAI. Workshop on Computational Aesthetics: AI Approaches to Beauty and Happiness*, Boston, MA, USA.

[Francisco et al., 2007] Francisco, V., Gervás, P., González, M., and León, C. (2007). EmoTag: Automated Mark Up of Affective Information in Texts. In *Proc. of AISB*, pages 179–186, Newcastle, UK.

[Francisco and Hervás, 2007] Francisco, V. and Hervás, R. (2007). EmoTag: Automated Mark Up of Affective Information in Texts. In Forascu, C., Postolache, O., Puscasu, G., and Vertan, C., editors, *Proc. of the Doctoral Consortium in EUROLAN 2007 Summer School*, pages 5–12, Iasi, Romania.

[García and Alías, 2008] García, D. and Alías, F. (2008). Emotion identification from text using semantic disambiguation. In *Procesamiento del Lenguaje Natural*, number 40, pages 75–82 *(in Spanish)*.

[Généreux and Evans, 2006] Généreux, M. and Evans, R. (2006). Towards a validated model for affective classification of texts. In *Sentiment and Subjectivity in Text, Workshop at the Annual Meeting of the Association of Computational Linguistics (ACL 2006)*, Sydney, Australia.

[Hofer et al., 2005] Hofer, G. O., Richmond, K., and Clark, R. A. J. (2005). Informed Blending of Databases for Emotional Speech Synthesis. In *Proc. Interspeech*.

[Iriondo et al., 2009] Iriondo, I., Planet, S., Socoró, J.-C., Martínez, E., Alías, F., and Monzo, C. (2009). Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification. *Speech Commun.*, 51(9):744–758.

[Islam and Inkpen, 2007] Islam, A. and Inkpen, D. (2007). Semantic Similarity of Short Texts. In *Proc. of RANLP*, Bulgaria.

[Joachims, 2006] Joachims, T. (2006). Training Linear SVMs in Linear Time. In *Proc. of KDD'06*, Philadelphia, PA, USA. ACM.

[Kim and Myaeng, 2007] Kim, Y. and Myaeng, S.-H. (2007). Opinion Analysis based on Lexical Clues and their Expansion. In *Proc. of NTCIR-6 Workshop Meeting*, pages 308–315, Tokyo, Japan.

[Koppel and Schler, 2006] Koppel, M. and Schler, J. (2006). The Importance of Neutral Examples for Learning Sentiment. *Comput. Intell.*, 22(2):100–109.

[Lan et al., 2009] Lan, M., Tan, C. L., Su, J., and Lu, Y. (2009). Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *IEEE T. Pattern. Anal.*, 31(4):721–735.

[Li and Ren, 2009] Li, H. and Ren, F. (2009). The study on text emotional orientation based on a three-dimensional emotion space model. pages 1–6, Dalian, China.

[Li and Zong, 2008] Li, S. and Zong, C. (2008). Multi-domain Sentiment Classification. In *Proc. of HLT-08*, pages 257–260, Columbus, Ohio. ACL, ACL.

[Liu, 2010] Liu, B. (2010). Sentiment Analysis and Subjectivity. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

[Manning and Klein, 2003] Manning, C. and Klein, D. (2003). Optimization, Maxent Models, and Conditional Estimation without Magic. In *Tutorial at HLT-NAACL 2003 and ACL 2003*, Edmonton, Canada.

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

[Mehrabian, 1995] Mehrabian, A. (1995). Framework for a comprehensive description and measurement of emotional states. *Genet. Soc. Gen. Psych.*, 121:339–361.

[Mohammad and Hirst, 2005] Mohammad, S. and Hirst, G. (2005). Distributional measures as proxies for semantic relatedness. In submission.

[Monzo et al., 2008] Monzo, C., Formiga, L., Adell, J., Iriondo, I., Alías, F., and Socoró, J. C. (2008). Adapting the URL-TTS for the Albayzin 2008 competition. In *Proc. of V Jornadas en Tecnología del Habla*, pages 87–90, Bilbao, Spain.

[Murray and Carenini, 2009] Murray, G. and Carenini, G. (2009). Detecting Subjectivity in Multiparty Speech. In *Proc. of Interspeech'09*, pages 2007–2010, Brighton, UK.

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.

[Osherenko, 2008] Osherenko, A. (2008). Towards Semantic Affect Sensing in Sentences. In *Proc. of AISB'08*, Aberdeen, Scotland, UK.

[Pang and Lee, 2005] Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proc. of the ACL*.

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proc. of EMNLP'02*, pages 79–86, Philadelphia, PA, USA.

[Plutchik, 1980] Plutchik, R. (1980). *Emotion: A Psychoevolutionary Synthesis*. Harper & Row, New York, NY, USA.

[Porter, 1980] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

[Read, 2004] Read, J. (2004). Recognising affect in text using pointwise-mutual information. Master's thesis, University of Sussex.

[Rebordao et al., 2009] Rebordao, A. R. F., Shaikh, M. A. M., Hirose, K., and Minematsu, N. (2009). How to Improve TTS Systems for Emotional Expressivity. In *Proc. of Interspeech'09*, pages 524–527, Brighton, UK.

[Reichel and Pfitzinger, 2006] Reichel, U. D. and Pfitzinger, H. R. (2006). Text Preprocessing for Speech Synthesis. In *Proc. of the TC-STAR Workshop on Speech-to-Speech Translation*, pages 207–212, Barcelona, Spain.

[Russell, 1980] Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.*, 39:1161–1178.

[Russell et al., 1989] Russell, J. A., Lewicka, M., and Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *J. Pers. Soc. Psychol.*, 57(5).

[Salton et al., 1975] Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

[Sassano, 2003] Sassano, M. (2003). Virtual Examples for Text Classification with Support Vector Machines. In *Proc. of EMNLP'03*, pages 208–215, Sapporo, Japan.

[Scherer, 1984] Scherer, K. R. (1984). Emotion as a multicomponent process: A model and some cross-cultural data. *Review of Personality and Social Psychology*, 5:37–63.

[Schröder, 2004a] Schröder, M. (2004a). Dimensional Emotion Representation as a Basis for Speech Synthesis with Non-Extreme Emotions. In *Proc. Workshop Affective Dialogue Syst.*, pages 209–220, Kloster Irsee, Germany. Springer LNAI.

[Schröder, 2004b] Schröder, M. (2004b). *Speech and Emotion Research: An overview of research frameworks and a dimensional approach to emotional speech synthesis*. PhD thesis, Research Report of the Institute of Phonetics, Saarland University.

[Schröder et al., 2001] Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S. (2001). Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis. In *Proc. of the 7th European Conference on Speech Communication and Technology (EUROSPEECH'01)*, pages 87–90, Aalborg. Kommunik Grafiske Losninger A/S.

[Schuller et al., 2009] Schuller, B., Steidl, S., and Batliner, A. (2009). The Interspeech 2009 Emotion Challenge. In *Proc. of Interspeech'09*, pages 312–315, Brighton, UK.

[Sebastiani and Ricerche, 2002] Sebastiani, F. and Ricerche, C. N. D. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47.

[Shaikh et al., 2008] Shaikh, M. A. M., Prendinger, H., and Ishizuka, M. (2008). Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. *Appl. Artif. Intell.*, 22(6):558–601.

[Shaver et al., 1987] Shaver, P., Schwartz, J., Kirson, D., and O'Connor, C. (1987). Emotion Knowledge: Further Exploration of a Prototype Approach. *J. Pers. Soc. Psychol.*, 52(6):1061–1086.

[Stevenson et al., 2007] Stevenson, R. A., Mikels, J. A., and James, T. W. (2007). Characterization of the Affective Norms for English Words by discrete emotional categories. *Behav. Res. Meth.*, 39(4):1020–1024.

[Strapparava and Mihalcea, 2007] Strapparava, C. and Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. In *Proc. of SemEval'07*, Prague, Czech Republic.

[Strapparava and Mihalcea, 2008] Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *Proc. of SAC'08*, pages 1556–1560, New York, NY, USA. ACM.

[Toutanova and Manning, 2000] Toutanova, K. and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proc. of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 63–70, Morristown, NJ, USA. ACL.

[Valitutti, 2004] Valitutti, A. (2004). WordNet-Affect: an Affective Extension of WordNet. In *Proc. of LREC'04*, pages 1083–1086, Lisbon, Portugal.

[Wang and Acero, 2007] Wang, Y.-Y. and Acero, A. (2007). Maximum entropy model parameterization with TF*IDF weighted vector space model. pages 213 –218.

[Watson and Tellegen, 1985] Watson, D. and Tellegen, A. (1985). Towards a consensual structure of mood. *Psychol. Bull.*, 98:219–235.

[Whissell, 2008] Whissell, C. (2008). A comparison of two lists providing emotional norms for English words (ANEW and the DAL). *Psychol. Rep.*, (102):597–600.

[Whissell, 1989] Whissell, C. M. (1989). The dictionary of affect in language. *Emotion: Theory, Research, and Experience*, pages 13–131.

[Wilson et al., 2009] Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Comput. Linguist.*, 35(3):399–433.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA, USA.

[Yu et al., 2009] Yu, D., Deng, L., and Acero, A. (2009). Using continuous features in the maximum entropy model. *Pattern Recogn. Lett.*, 30(14):1295–1300.