Analysis of Inconsistencies in Cross-Lingual Automatic ToBI Tonal Accent Labeling

David Escudero-Mancebo, Carlos Vivaracho Pascual, César González Ferreras, Valentín Cardeñoso-Payo[1], and Lourdes Aguilar[2]

[1] Dpt. of Computer Sciences, Universidad de Valladolid, Spain
[2] Dpt. of Dpt. of Spanish Philology, Universidad Autónoma de Barcelona, Spain

Abstract. This paper presents an experimental study on how corpus-based automatic prosodic information labeling can be transferred from a source language to a different target language. Tone accent identification models trained for Spanish, using the ESMA corpus, are used to automatically assign tonal accent ToBI labels on the (English) Boston Radio news corpus, and vice versa. Using just local raw prosodic acoustic features, we got about 75% correct annotation rates, which provides a good starting point to speed up automatic prosodic labeling of new unlabeled corpora. Despite the different ranges and relevance of inter corpora acoustic input features, the contrasting of the results with respect to manual labeling profiles indicate the potential capabilities of the procedure.

Index Terms: prosodic labeling, cross-lingual prosody, automatic accent identification.

1 Introduction

ToBI has been implemented for several languages including English, German and Japanese. Despite the intensive research activity for Iberian languages, the need of a reference corpus similar to those existing for other languages (e.g. the Boston Radio Corpus for English [9]) is still a need both for Catalan and Spanish. The activity presented in this paper is included in the Glissando project[1], which aims to record and label a bilingual Spanish and Catalan corpus that contains Radio news recordings and spontaneous dialogs with ToBI marks [11]

Labeling a corpus with ToBI tags is an expensive procedure. In [12] it is estimated that the ToBI labeling commonly takes from 100-200 times the real time. To speed up the process, automatic or semiautomatic methods would seem to be a productive resource. [2] or [10] are good examples of the state of the art on automatic labeling of ToBI events. For Catalan, [1] presents a procedure to label break indices, reducing the set of break indices by merging some of them together in order to increase the identification results. This merging strategy is common in other studies such as the ones already mentioned from [2] or [10] that combine the different types of accent tones, transforming the labeling problem into a binary one in order to decide whether

[1] Partially funded by the Ministerio de Ciencia e Innovacion, Spanish Government Glissando project FFI2008-04982-C003-02 and Junta de Castilla y León JCYL2011.

I. Habernal and V. Matoušek (Eds.): TSD 2011, LNAI 6836, pp. 41–48, 2011.

Table 1. Number of words in the corpora and subcorpora

	Words			Syllables		
	#total	#accented	#un-accented	#total	#accented	#un-accented
ESMA-UPC	7236	2483	4753	14963	2341	12622
BURNC	27767	13899	13868	47323	14873	32450

an accent is present or not. In this paper an automatic labeling procedure of accent tones (binary decision) is presented that aims to speed up the job of the manual labelers who will be required to check the predictions of the system.

Here we explore a cross-lingual approach where a given corpus with ToBI labels will be used to predict the labels of a different corpus in a different language. Despite the fact that the ToBI sequences are highly dependent on the language, they codify universal functions of prosody, including the prominence (here prominence is identified with the presence of a ToBI accent tone mark). Thus, the Boston Radio Corpus is used to train prosodic models that are then used to identify the presence of tonal accent in a Spanish corpus and vice versa. Results are promising as, using raw prosodic features, close to 75% of the words are correctly classified.

This cross-lingual approach is thus an opportunity for prosodic studies, as the number of linguistic resources with ToBI labels is sparse and the number of languages that lack this information is still large. At the same time, we assume that there are challenges to cope with, so differences in the relevance and scale of the input features in the different languages are analyzed and identified as the battlefield on wich to increase the performance of the classifier in future works. First, the experimental procedure is presented followed by the results on cross-lingual accent identification. Discussion and future work end the paper.

2 Experimental Setup

The cross-lingual approach does not solely consist of training classification models with the data of a corpus in a given language in order to test with the data of a different one. Additionally, we systematically contrast the differences on the input features among the diverse corpora in several practical aspects. First, the scale of the input features is analyzed to contrast the differences among the languages and the impact of the normalization of the input is shown.

The cross-lingual study continues with the examination of the relevance of the input features in the different corpora. Input features are ranked in terms of theirs informative capabilities for discriminating whether a word or a syllable is accented. Every language has their own ranking to be contrasted. Furthermore, the most informative input features are also analyzed to verify whether significant differences appear among the diverse corpora (e.g. f0_range of the accents in Spanish vs. accents in English).

Finally, automatic prediction results are contrasted with perceptual judgements made by a set of labelers on the same testing corpus. This is useful to value the usefulness of the automatic labeling process in general and of the cross-lingual labeling in particular.

Table 2. Classification rates (in percentages) using words in terms of the presence of accent. The training/testing corpora in the rows. C4.5 is the decision tree; MLP is the neural network refering to each type of classifier used. *NI* in normalized input. *Ov* is oversampled input. *T* is total, *A* is accented and *U* is unaccented.

Corpus	C4.5($\bar{NI}$)($\bar{Ov}$)			C4.5(NI)($\bar{Ov}$)			C4.5 ($\bar{NI}$)(Ov)			C4.5(NI)(Ov)			MLP(NI)(Ov)		
Trainning/Testing	T	A	U	T	A	U	T	A	U	T	A	U	T	A	U
ESMA-UPC/ESMA-UPC	79.6	64.0	87.7	79.6	64.0	87.7	79.6	81.2	77.8	78.7	81.9	75.3	77.0	75.7	77.6
ESMA-UPC/BURNC	65.1	36.5	93.8	71.2	55.4	87.1	68.3	46.4	90.3	73.4	67.8	79.0	72.6	67.5	77.8
BURNC/ESMA-UPC	61.8	84.1	50.2	73.2	78.4	70.5	68.8	86.5	50.2	74.8	78.9	70.5	66.6	81.1	59.0
BURNC/BURNC	77.0	75.6	78.4	77.6	78.2	77.1	77.0	75.6	78.4	77.6	78.2	77.1	78.8	80.1	77.5

2.1 Processing of the Corpora

We used the Boston University Radio News Corpus BURNC[9]. This corpus includes labels separating phonemes, syllables and words. Accents are marked with a ToBI label and a position. Inspired in state of the art works [14,2], the accent tones were aligned with respect to the prominent syllable and the word containing it (table 1).

The Spanish corpus used in this paper is ESMA-UPC. It was designed for the construction of a unit concatenative TTS system for Spanish [3]. Although it was not specifically designed for prosodic studies, it contains enough data to get significant results. Speech was recorded in one channel and the output of a laryngograph in another one. Data were automatically labeled and manually supervised. Labeling included silences, allophonic transcription, and allophonic boundaries. This information was increased by the additional syllable and word boundaries and accent positions(table 1).

Feature extraction in both corpora was carried out using similar features to other experiments reported in the bibliography [2]. The features concern frequency: within word F0 range (f0_range), difference between maximum and average within word F0 (f0_maxavg_diff), difference between average and minimum within word F0 (f0_minavg_diff), difference between within word F0 average and utterance average F0 (f0_avgutt_diff); energy: within word energy range (e_range), difference between maximum and average within word energy (e_maxavg_range), difference between average and minimum within word energy (e_minavg_range); duration: maximum normalized vowel nucleus duration from all the vowels of the word (normalization is done for each vowel type) (duration).

Although POS Tagging information and context have shown to be helpful in the improvement of the classification results (see [14,2]), this information was not used in the experiments reported in this paper. There is no obvious correspondence between POS tags used in each corpus: BURNC corpus uses the Penn Treebank tag set[8] and ESMA uses the EAGLES tag set for Spanish labeled using the Freeling tagger (`http://www.freeling.org`). A valid correspondence is under study for its application in future works. Regarding to context, we focus on local effects (at the level of word and/or syllable) as the context can be highly dependent on the language and the modeling of its correspondence is beyond the scope of this paper.

2.2 The Classifiers

A Multilayer Perceptron (MLP) with a non-linear sigmoid unit is trained for each classification problem, using the Error Backpropagation learning algorithm. Several network

configurations were tested, achieving the best results with the following: i) single hidden layer with 12 neurons (more hidden units than inputs were used to achieve separation surfaces between closed classes), ii) 100 training epochs, iii) two neurons in the output layer, one for each class to be classified, then the test input vector is assigned to the class corresponding to the largest output.

The Weka machine learning toolkit [7] was used to build C4.5 decision trees (J48 in Weka). Different values for the confidence threshold for pruning have been tested, although the best results are obtained with the default value (0.25). The minimum number of instances per leaf is also set to the default value (2).

In [6] the negative impact of imbalanced data on final result is shown. Therefore, re-sampling methods are applied: minority class example repetition [13] for the MLP classifier and Synthetic Minority Oversampling TEchnique (SMOTE) method [4] for the C4.5 classifier. Due to the different scale of the features among the training corpora, we applied different normalization techniques: the Z-Norm, Min-Max, divide by maximum and euclidean norm 1. The normalization has been processed by corpus and by speacker.

3 Results

In spite of the fact that input features are relative magnitudes (differences with respect to a mean value), significant differences appear between the diverse corpora. These differences were expected, independently of the cross-lingual effect, as the different recording conditions have a clear impact on the values of the input magnitudes. Thus, for example, the ESMA F0 values have been collected with a laringograph device and BURNC F0 values with a pitch tracking algorithm leading to less stable values. Differences between the Spanish corpus and the English one are clear, so that the normalization of the input features is thus a need in this work.

The impact of the normalization is clearly seen in table 2 (columns NI vs. $\bar{NI}$): results significantly improve when the input is normalized in the cross-corpus scenarios. Oversampling the corpora (Ov vs $\bar{Ov}$ columns) has also a positive impact to reduce the imbalanced results corresponding the accented vs. unaccented classes. In the conventional scenarios (same corpus for training and testing) results are obtained going up to 79.6%, which are the expected results according to the state of the art: [10] reports state of the art up-to-date results from 75.0% to 87.7% using the Boston Radio Corpus but adding the morpho-syntactic POS tag information and taking into account the context. In the cross-lingual scenarios, the classification rates decrease but they get to acceptable results taking into account that a posterior manual revision of the predictions will be applied. Syllables are also used as the reference unit with similar results (words in tables 2).

In order to analyze the reasons for the performance decrease in the cross-lingual accent classification task, we contrast the informative capabilities of the input features in the Spanish and English corpora. Table 3 compares the *Information Gain* of the different features, providing a measure of the potential loss of entropy which would be generated if the splitting of the training set was carried out in terms of the present feature [15]. The tagging of the Spanish corpus seems to rely mainly on F0 features, as

Table 3. Info Gain (IG), computed with the WEKA software, of the features when they are used to classify the accents in the different corpora

ESMA-UPC		BURNC	
Feature	IG	Feature	IG
f0_minavg_diff	0.18888	f0_minavg_diff	0.248
f0_range	0.18246	f0_range	0.231
f0_avgutt_diff	0.15215	f0_maxavg_diff	0.191
f0_maxavg_diff	0.10891	e_range	0.184
e_range	0.09695	e_maxavg_diff	0.162
e_minavg_diff	0.08156	duration	0.159
e_maxavg_diff	0.07681	e_minavg_diff	0.141
duration	0.0063	f0_avgutt_diff	0.121

Table 4. Statistics of the input feature f0_minavg_diff

	ESMA-UPC	BURNC
Accented data:	mean=0.67 sd=1.00	mean=0.48 sd=1.07
Unaccented data:	mean=-0.35 sd=0.80	mean=-0.49 sd=0.62
Total:	mean=0.0 sd=1.0	mean=0.0 sd=1.0

the four most relevant features are related with F0 and the difference with respect to the energy and duration features is important. The tagging of the English corpus also seems to rely mainly on F0 features ($f0\_minavg\_diff$ and $f0\_range$ share the top ranking position in both corpora). Nevertheless, differences appear as the energy and duration appear to be more relevant for the English transcribers than for the Spanish ones.

Table 4 shows the discrimination capabilities of the most informative input feature (f0_minavg_diff) for the ESMA and the BURNC corpora. There are no statistical differences for this variable between the Spanish and English corpora (see the row that has the title *Total* in the table 4). Nevertheless, when the data are split into the classes, statistically significant differences clearly appear between the accented words of the diverse corpora and between the unaccented wordds (t-test p-value=2.2e-16). These differences justify the performance decrease of the classification task in cross-lingual situations observed in table 2.

4 Contrasting Automatic and Manual Labeling

One of the advantages of using automatic classifiers in contrast with manual labeling is the introduction of objective criteria. In [5] an experiment was presented where a subset of 20 utterances from the ESMA corpus were manually labeled by a team of six ToBI human raters. The inter-transcriber agreement indicated that there are two groups of labelers using two potential labeling criteria. In group one the inter-pair agreement goes from 86.7 to 92.0% while agreement of the 2 labelers of the second group is 94.7%. The inter group agreement decreases, going down to the range 65.5 to 73.5%. We hypothesize here that the availability of automatic prediction is an opportunity not only to speed up the manual labeling process, but also to offer an objective criteria to

```
F0_RANGE <= 0.252894: F (4567.0/839.0)                                    ESMA
f0_range > 0.252894
|   f0_avgutt_diff <= -0.33511: F (564.0/119.0)
|   f0_avgutt_diff > -0.33511
|   |   e_range <= -0.780483
|   |   |   duration <= 1.512244: F (107.0/30.0)
|   |   |   duration > 1.512244: T (15.0/1.0)
|   |   e_range > -0.780483: T (1983.0/502.0)
```

Fig. 1. Decision tree C4.5. Simplified version with pruning confidence of 0.001 (default is 0.25). T in the branches is accented word and F is unaccented word.

Table 5. Most frequent consistencies and inconsistencies. Rows refer to the corpus used to train the classifier. Columns refer to the groups of labelers to contrast with. In the cells the information is A/B(C), being A the label set by the human transcriber, B the label assigned by the automatic classifier and C is the percentage of observations among all the inconsistencies (upper table) or among all the consistencies (bottom table).

Most common disagreements

	Group 1 of labelers	Group 2 of labelers
ESMA-UPC	L*/noAccent (35.1) H*/noAccent (12.4) L+H*/noAccent (11.4) L+>H*/noAccent (10.9)	L*/noAccent (35.3) 0/Accent (24.3) L+H*/noAccent (13.2) L+>H*/noAccent (11.8)
BURNC	L*/noAccent (36.0) !H*/noAccent (10.3) L+H*/noAccent (9.8) H*/noAccent (9.8)	L*/noAccent (40.3) 0/Accent (19.4) L+>H*/noAccent (10.9) L+H*/noAccent (10.1)

Most common agreements

	Group 1 of labelers	Group 2 of labelers
ESMA-UPC	L+>H*/Accent (26.0) 0/noAccent (19.2) L+H*/Accent (12.4) L*/Accent (12.0)	0/noAccent (35.0) L+>H*/Accent (25.6) L+H*/Accent (12.3) L*/Accent (10.3)
BURNC	L+>H*/Accent (28.6) 0/noAccent (20.2) L+H*/Accent (13.9) H*/Accent (12.6)	0/noAccent (37.6) L+>H*/Accent (25.7) L+H*/Accent (14.3) L*/Accent (8.1)

help to improve the inter-rater agreement or to resolve labeling inconsistencies. In this example, the automatic predictions are clearly closer to the second group of labelers with differences of 18.2 points average using the ESMA-UPC as the training corpus and 12.8 using the BURNC corpus, an objective indicator being to select a labeler or the labels from one of the two groups.

This mentioned objective criterion can be defended by using the resulting decision trees (see figure 1). The essential interpretation of the decision trees has been observed to be the same for both languages: the *accented* tag is assigned to combinations of features with a high variability of F0 and/or energy and/or long duration. The cross-lingual observed differences affects to the position of the features in the levels of the trees and to the threshold values used in the binary decisions. As the classification results are contrasted, a common inter classifier behavior is observed. Thus for example, table 5 shows that the most frequent disagreement is the same both for the ESMA-UPC trained classifier and the BURNC trained one: A high number of L* tonal accents are classified

as *unaccented* which represents more than 35% of all the disagreements in both cases. Furthermore, the four most common disagreements, representing more than 80% of the total amount of disagreements with respect to the symbols assigned by the group two of labelers, are shared by both classifiers. Again, the four most common agreements representing more than the 80% of the agreements are also the same for both classifiers. This result evidences a similar behavior of the classifiers, and encourages for using cross-lingual labeling of prosodic event in combination with a posterior supervised revision of the results by human labelers in future works.

5 Conclusions and Future Work

A cross-lingual experiment on tonal accent identification has been presented. The two corpora used have been presented and the experimental strategy has been described. Results indicate that the automatic classifiers offer an objective criterion that permits close to 75% of the input units in cross lingual situations to be identified. The inter-corpus input features scale differences force the revision of the results by the manual labelers. Nevertheless, the predictions of the classifiers can be considered as an objective reference to speed-up the ToBI labeling of the target corpus.

Cross-lingual differences appear in the relevance of the input features and their distribution of values with negative impact in the labeling results. To overcome this difficulty is the challenge for future work with the use of speaker adaptation techniques, introduction of more representative input features or the inclusion of language dependent information in the normalization process.

References

1. Aguilar, L., Bonafonte, A., Campillo, F., Escudero, D.: Determining Intonational Boundaries from the Acoustic Signal. In: Proceedings of Interspeech 2009, pp. 2447–2450 (2009)
2. Ananthakrishnan, S., Narayanan, S.: Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. IEEE Transactions on Audio, Speech, and Language Processing 16(1), 216–228 (2008)
3. Bonafonte, A., Moreno, A.: Documentation of the upc-esma spanish database. Tech. rep., TALP Research Center, Universitat Politecnica de Catalunya, Barcelona, Spain (2008)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
5. Escudero, D., Aguilar, L.: Procedure for assessing the reliability of prosodic judgements using Sp-TOBI labeling system. In: Proceedings of Prosody 2010 (2010)
6. Gonzalez, C., Vivaracho, C., Escudero, D., Cardenoso, V.: On the Automatic ToBI Accent Type Identification from Data. In: Interspeech 2010 (2010)
7. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1), 10–18 (2009)
8. Meteer, M., Schwartz, R.M., Weischedel, R.M.: Post: Using probabilities in language processing. In: IJCAI, pp. 960–965 (1991)
9. Ostendorf, M., Price, P., Shattuck, S.: The boston university radio news corpus. Tech. rep., Boston University (1995)

10. Rangarajan Sridhar, V., Bangalore, S., Narayanan, S.: Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework. IEEE Transactions on Audio, Speech, and Language Processing 16(4), 797–811 (2008)
11. Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J.: ToBI: A standard for labelling English prosody. In: Proceedings of ICSLP-1992, pp. 867–870 (1992)
12. Syrdal, A.K., Hirshberg, J., McGory, J., Beckman, M.: Automatic ToBI prediction and alignment to speed manual labeling of prosody. Speech Communication (33), 135–151 (2001)
13. Vivaracho-Pascual, Simon-Hurtado, A.: Improving ann performance for imbalanced data sets by means of the ntil technique. In: IEEE International Joint Conference on Neural Networks (July 18-23, 2010)
14. Wightman, C., Ostendorf, M.: Automatic labeling of prosodic patterns. IEEE Transactions on Speech and Audio Processing 2(4), 469–481 (1994)
15. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (1999)