Cross-Language Prominence Detection

Andrew Rosenberg[1], *Erica Cooper*[2], *Rivka Levitan*[2], *Julia Hirschberg*[2]

[1]Department of Computer Science, Queens College / CUNY, USA
[2]Department of Computer Science, Columbia University, USA
andrew@cs.qc.cuny.edu, {ecooper, rlevitan, julia}@cs.columbia.edu

Abstract

We explore the ability to perform automatic prosodic analysis in one language using models trained on another. If we are successful, we should be able to identify prosodic elements in a language for which little or no prosodically labeled training data is available, using models trained on a language for which such training data exists. Given the laborious nature of manual prosodic annotation, such a process would vastly improve our ability to identify prosodic events in many languages and therefore to make use of such information in downstream processing tasks. The task we address here is the detection of intonational prominence, performing experiments using material from four languages: American English, Italian, French and German.

While we do find that cross-language prominence detection is possible, we also find significant language-dependent differences. While we hypothesized that language family might serve as a reliable predictor of cross-language prosodic event detection accuracy, in our experiments this did not prove to be the case. Based upon our results, we suggest some directions that may be able to improve our cross-language approach.

Index Terms: prosody, prominence, language independence

1. Introduction

Detecting prosodic events in speech has been shown to be useful for automatic corpus annotation for part-of-speech tagging and syntactic disambiguation, automatic annotation of corpora for text-to-speech synthesis, reducing language model perplexity for speech recognition, salience detection and distinguishing between given and new information in speech summarization, identifying turn-taking behavior and dialogue acts in spoken dialogue systems, e.g. [1, 2, 3, 4, 5, 6]. However, training prosodic event models requires large amounts of hand-labeled training data, which is not available for most languages. Our goal is to determine whether prosodic models trained on labeled data in one language can be adapted succesfully to identify prosodic events in another language for which minimal, if any, labeled data exists. We experiment with detecting intonational prominence, or **pitch accent**, using labeled data from Standard American English (SAE), French, German, and Italian.

There is a wealth of research in the automatic detection of prominence in, for example, SAE [7], French [8], German [9], and Italian [10]. However, there is very little work that takes a comparative view to prominence across languages. In one exception, Tamburini [10] explored how parameter settings in a model of prominence prediction could be configured to better model Italian, Dutch and SAE prominence. Maier et al. [11] proposed a “language-independent” feature set for prosodic analysis. Other work has examined the perception and production of L2 prosody by L1 speakers, e.g. [12]. To our knowledge there are no published experiments that have explored the cross-language adaptation of prosodic analysis models.

In Section 2 we describe the corpora we use in our experiments. In Section 3 we describe our approach to prominence detection from labeled corpora. In Section 4, we describe our cross-language adaptation and language-independent prediction experiments. In Section 5 we compare the value of different features for predicting prominence in each language, and in Section 6 we compare characteristics of the four languages to explain our results. Section 7 presents semi-supervised domain adaptation experiments leveraging English training data to improve models for other languages.

2. Material

We examined four corpora in four different languages for our experiments: the Boston Directions Corpus, the DIRNDL Corpus, the C-PROM Corpus, and Italian read speech. We note that the durations reported for each corpus include only words, ignoring silences.

Boston Directions Corpus (BDC) – The BDC [13] is comprised of both read and spontaneous monologues elicited from four non-professional speakers, three male and one female. Speakers were asked to perform increasingly complex direction-giving tasks. Their directions were recorded and transcribed. Several weeks later, the subjects returned and were recorded reading transcriptions of their own directions. There are approximately 60 minutes of read speech and 50 minutes of spontaneous speech. The corpus is orthographically transcribed and ToBI-labelled. We use three speakers for training material and the fourth (h2) as test material. The training data has 13,975 words (55.85 mins) with an accent rate of 46.75%; the test data contains 8,483 words (32 mins) with an accent rate of 42.14%.

C-PROM – Our French data comes from the C-PROM corpus [14], a corpus annotated for French prominence studies. The corpus includes about 70 minutes of speech from 28 speakers, 12 female and 16 male. It contains a mix of seven genres ranging in formality from life story to radio interview to read speech. We divide the C-PROM corpus into training and testing splits with no speaker overlap, using the NAR (Life Stories) and POL (Political Speeches) genres for testing. The training data has 9,387 words comprising 40.33 minutes with an accent rate of 29.74%; the test data contains 3,794 words (15.75 minutes) with an accent rate of 27.10%.

DIRNDL – The Discourse Information Radio News Database for Linguistic analysis corpus [15] is a database of German radio news broadcasts. It contains approximately two and a half hours of radio news, along with accompanying transcripts from which fillers, disfluencies and music have been removed. The corpus is annotated for intonation according to GToBI. We divide the DIRNDL material into training and testing splits with no speaker overlap. The training data has 9200 words (57.98 mins) with an accent rate of 52.65%; the test data has 3695 words (28.78 mins) with an accent rate of 49.15%.

Italian – Our Italian corpus contains about 25 minutes of read speech from a single male professional speaker; this corpus was made available to us by Cinzia Avesani at the Institute of Cognitive Sciences and Technologies in Padova. The speaker reads two different short stories. The corpus is orthographically transcribed and prosodically annotated for Italian ToBI. As this corpus contains material from a single speaker, the Italian experiments are speaker-*dependent* in contrast to the other corpora. The training data contains 2,780 words (18.55 mins) with an accent rate of 57.34%; the test data contains 1,095 words (7.32 mins) with an accent rate of 55.42%.

3. Prominence Detection

To identify intonational prominences, we employ the AuToBI [16] system. AuToBI uses an L2-regularized logistic regression classifier for prominence detection. This tool was developed to identify prosodic events (pitch accents and prosodic phrase boundaries) in Standard American English (SAE).

The features AuToBI extracts and uses in prosodic event prediction can be divided into four feature types: **Pitch** features are extracted from the raw and speaker normalized pitch contour using log Hz as the unit of pitch. Speaker normalization is performed by z-score normalization based on statistics collected from each audio file. The minimum, maximum, mean, standard deviation and z-score of maximum of the pitch contour and its slope within a word are included in the feature vector. We also extract the mean and maximum pitch normalized by the surrounding acoustic context. This context is defined by all combinations of up to 2 previous and 2 following words, leading to 8 context windows. **Intensity** features are extracted in a similar manner to the pitch features. Intensity contours are represented in decibels. **Spectral balance** features are also extracted similarly to those extracted from pitch and intensity contours. They are extracted from a contour of spectral balance extracted at 10ms frame, and calculated as the ratio of the energy in the speech signal between 2 and 20 bark to the total energy in the frame. AuToBI also extracts **pause/duration** features based on the length of the current word and the length of preceding and following silence. The decision to predict prominence at the word, rather than the syllable level is, due to previous experimental findings in SAE [7]. The value of context normalization is also experimentally supported [7, 17].

4. Cross-language Prominence Detection

In this section we explore the use of prominence detection models trained on material from one language and evaluated on another. This can be viewed as a domain transfer experiment, in which models from one domain (source language) are applied to another. We first explore the hypothesis that language family may help to predict accuracy of cross-language prominence detection. That models trained on Romance languages like French might perform better on other Romance languages, like Italian, and vice versa, and that Germanic languages like English would perform better on German than on Romance languages, and vice versa.

We train these models in two ways, using either the full corpus as training data, or using only a randomly selected 25 minutes of training data. The corpus size is chosen as the size of our smallest corpus, the Italian corpus, which contains 25 minutes of speech data, so that all of our models can be trained on the same amount of material. We evaluate the models using only the testing split from each corpus as evaluation data. The results from the experiments using models trained on the full corpora can be found in Table 1. The results from experiments using models trained on 25 minutes of material can be found in Table 2. The hypothesis that models trained on a language from the same language family might perform better than models trained on different families is not supported by these experiments. Instead, we find that the German material is the most reliable model for prediction of prominence in other languages. The language family hypothesis suggests that English and German should generate compatible models, as should French and Italian. However, we find that DIRNDL (German) models predict C-PROM (French) and Italian prominence better than models trained on their within-language-family counterparts. Moreover, models trained on Italian or French material predict German prominence better than they predict prominence in any other language.

	Training Corpus – Full			
	BDC	C-PROM	DIRNDL	Italian
BDC	-	71.99(0.34)	76.28(0.44)	62.95(0.12)
C-PROM	80.35(0.28)	-	84.29(0.42)	79.28(0.24)
DIRNDL	76.97(0.53)	82.90(0.65)	-	82.08(0.64)
Italian	80.22(0.56)	77.20(0.49)	80.95(0.57)	-

Table 1: *Accuracy and (in parentheses) relative error reduction using models trained on full corpora in one language and testing on another.*

	Training Corpus – 25 mins			
	BDC	C-PROM	DIRNDL	Italian
BDC	-	71.88(0.33)	78.07(0.48)	62.95(0.12)
C-PROM	79.44(0.24)	-	83.50(0.39)	79.28(0.24)
DIRNDL	78.43(0.55)	82.27(0.64)	-	82.08(0.64)
Italian	80.40(0.56)	78.40(0.52)	80.95(0.57)	-

Table 2: *Accuracy and (in parentheses) relative error reduction using models trained on 25 minutes of material from one language and testing on another.*

We now explore the possibility of language-independent prominence prediction, in which we train a model using material on three languages and evaluate the performance on the fourth. This approximates the process of training a single prominence detection system based on a diverse set of training material drawn from many languages and using it to predict prominence on an unknown language. We perform these evaluations using both the test split and the full corpus for each language. Results using the test-split are reported in Table 3.

Test Corpus	Accuracy	Majority Class Baseline
BDC	74.86%	57.86
C-PROM	81.81%	72.9
DIRNDL	84.24%	55.42
Italian	84.69%	50.85

Table 3: *Prominence detection accuracy training on three languages and testing on the test-split of the fourth.*

We find that the performance on the Italian and DIRNDL material is improved over the single language cross-language experiments, while for BDC and C-PROM the inclusion of training material other than DIRNDL leads to reduced performance. This suggests that for BDC and C-PROM, a model *selection* approach that is able to identify the DIRNDL model to be the most compatible would be the best language independent approach to prominence detection. However, training on a diverse range of language and applying the model to a previously unseen language is a more appropriate strategy for the Italian and DIRNDL material. We next consider how one might determine how to choose the best procedure for future languages.

5. Within-language Feature Analysis

In Section 3, we described four feature types that are used to predict prominence by AuToBI. Here we explore the relative performance of each of these within each language. These experiments allow us to compare the prosody of the four investigated languages from an acoustic perspective. We hypothesize that, if two languages are demonstrating similar relative prominence prediction performance using the four feature sets, they may be more compatible for cross-language prediction.

We train and evaluate models using training and testing splits from the same corpora. It is worth reiterating that the Italian training/testing splits contain material from the same speaker, while the C-PROM, DIRNDL and BDC material contains distinct speakers in the train and test sets. Results from these experiments can be found in Table 4. A chart of the relative reduction of error from a majority class baseline is presented in Figure 1. We see that each language makes use of acoustic correlates of prominence in distinct ways. On the Italian and German material, we see a similar error reduction profile from the four feature sets; duration is the most predictive feature set, with pitch, intensity and pause/duration features all providing a relatively similar reduction of error. On the English material (BDC) we find intensity and pause/duration to provide similar high reduction of error with pitch and spectral features to provide somewhat less. However, on the French (C-PROM) material, the pitch features offer the greatest reduction of error, somewhat more than intensity and pause/duration features, with spectral features providing only 14.9% error reduction.

Corpus	All	Pitch	Intensity	Duration	Spectral
BDC	79.55	71.68	75.61	77.00	72.45
C-PROM	86.11	82.63	80.73	81.26	76.94
DIRNDL	84.65	75.72	76.13	83.84	73.02
Italian	86.51	77.85	79.22	87.51	78.49

Table 4: *Accuracy using feature subsets.*

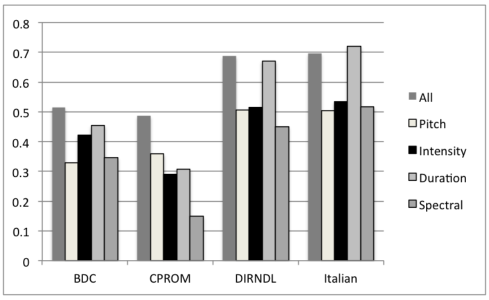

Figure 1: *Relative reduction of error using feature subsets.*

While we see clear similarities and differences between the four languages through this analysis, these relationships are not predictive of the cross-language results we observe in Section 4. Specifically, these results suggest that the Italian and DIRNDL corpora should be mutually compatible, while the BDC and CPROM material should yield lower accuracy. However, we find that the BDC material can predict Italian prominence nearly as well as DIRNDL, and that C-PROM-trained models predict DIRNDL prominence as well as Italian-trained models.

6. Comparing Feature Distributions

In Section 5, we find that the relative discriminative power of different feature sets did **not** predict how well each cross-language model would work. We continue to look for an explanation of why the prominence detection models based on the DIRNDL corpus predicts prominence in the other three languages better than any other models.

We compare the distribution of values of four representative features drawn from the four feature sets identified in Section 3. For **pitch**, **intensity** and **spectral balance**, we examine the speaker-normalized mean value context normalized by one previous and one following word. For the **pause/duration** features, we examine the duration of the current word. In Table 5, we present the mean and standard deviation of each value for prominent and non-prominent tokens in the training data of each language. We can treat these means and standard deviations as components of a Gaussian mixture model (GMM) with two components – one for prominent and the other for non-prominent tokens. Using these GMMs to describe the prosody of each language, we can measure the KL-divergence between each pair of languages for each feature. As no closed-form expression exists for the KL-divergence of GMMs, we use a Monte Carlo estimation with 100,000 samples to calculate these values [18]. To describe the similarity between a pair of languages we sum the four KL-divergence values based on each feature set extracted over the full corpus. Lower KL-divergence values indicate that the two distributions are more similar. We hypothesize that more similar distributions predict that languages will be more compatible for cross-language domain transfer. The total KL-divergence between each pair of languages is shown in Table 6. These results again indicate how similar the Italian and DIRNDL corpora are with respect to prominence. The similarity between the Italian and C-PROM material may explain why the cross-language results on these two corpora are nearly as good as the Italian/DIRNDL results. However, these results do not explain why the DIRNDL-trained model is able to predict prominence in BDC as well as it does.

Corpus	feature	prom.	non-prom.
BDC	pitch	0.166±0.85	-0.134±1.02
	int.	0.079±0.41	-0.155±0.63
	spec.	0.133±0.46	-0.211±0.41
	dur.	0.328±0.15	0.159±0.10
C-PROM	pitch	0.287±0.64	-0.204±0.80
	int.	0.075±0.41	-0.064±0.65
	spec.	0.098±0.43	-0.066±0.53
	dur.	0.425±0.19	0.186±0.14
DIRNDL	pitch	0.075±0.57	-0.216±0.79
	int.	0.053±0.32	-0.121±0.56
	spec.	0.027±0.31	-0.079±0.42
	dur.	0.529±0.22	0.230±0.13
Italian	pitch	0.032±0.71	-0.025±0.86
	int.	0.017±0.34	-0.038±0.66
	spec.	3.568±0.35	-0.032±0.49
	dur.	0.561±0.23	0.190±0.13

Table 5: *Mean and std. dev. of example features from the four feature sets*

Corpus	BDC	C-PROM	DIRNDL	Italian
BDC	0	0.126	0.493	0.402
C-PROM	-	0	0.375	0.266
DIRNDL	-	-	0	0.056
Italian	-	-	-	0

Table 6: *Total KL divergence between each pair of languages based on supervised GMM based on four features.*

7. Adaptation with Augmented Data

There is more publicly available prosodically annotated material in SAE than any other language. This is reflected here insofar as the BDC corpus is our largest corpus.

In this experiment we explore a semi-supervised, domain adaptation technique to leverage the amount of training data available in English to improve models from other languages. We begin with a base model trained on the full BDC corpus. We then augment the training data with increasing amounts of labeled training data from the target language. The training data in the target language is drawn from the training split and the evaluation data is the test split as defined in Section 2. Prominence detection accuracy with increasing amounts of training data is shown in Figure 2. We compare prominence prediction accuracy on the test partitions using models trained on a limited amount of within-language training data, and, optionally, all of the BDC training material. This approximates a situation where there is very little available training data in a low resource language. We address the questions of whether it is better to train a model based on a small amount of in-language training data or to augment this training data using a semi-supervised approach.

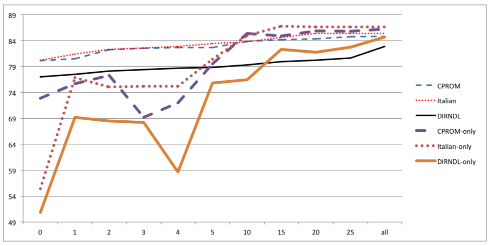

Figure 2: *Accuracy from models trained on BDC material augmented with variable amounts of target-language training data.*

In these experiments, we find that the inclusion of any training data in the target domain leads to improvements to target language performance. We also find that when there is less than 10 minutes of available training data in the target language, better performance is obtained by data augmentation from a larger language corpus. However, when 15 minutes of annotated data is available, language-specific models generate greater accuracy, though not to a statistically significant degree under a proportion test. We also find that the data-augmentation models demonstrate monotonic increase as within-language training data is added to the BDC material. This result suggests that this is a relatively stable approach to adapting models from one language to another. The rightmost column represents the inclusion of all of the available training data. There is no language where this data augmentation style of domain transfer improves prominence detection performance over the within-language performance given all of the available training data. However, when there is *very* little annotated training data – less than ten minutes – available for a language, augmenting English training material can serve to improve performance.

8. Conclusion and Future work

This paper describes experiments in cross-language prosodic event detection. We have found that there are differences in the cross-language performance across languages. However, language family does not appear to be predictive of cross-language prominence detection, at least in our experiments. Nor is the relative importance of features used in prominence prediction. However, we do find that augmented training data can be used to adapt American English models to other languages, leading to significant performance improvements. Also, including training material from a variety of source languages can, for some languages, improve performance over that observed using training data from a single language. In future research, we will further explore domain adaptation for prosodic event detection. In developing ideas for this work, we investigated the "frustratingly easy" adaptation approach described in [19], but found it to yield frustratingly poor results for this adaptation task. We will explore other supervised and unsupervised adaptation techniques to facilitate cross-language adaptation of prosodic analysis modules. These explorations will also include other prosodic analysis tasks including prominence type classification, phrasing detection and phrase-ending classification.

9. References

[1] Z. Huang and M. Harper, "Appropriately handled prosodic breaks help pcfg parsing," in *HLT-NAACL* 2010.

[2] V. Eidelman, et al., "Lessons learned in part-of-speech tagging of conversational speech," in *EMNLP* 2010.

[3] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S. Kim, J. Cole, and J. Choi, "Prosody dependent speech recognition on radio news," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 1, pp. 232–245, 2006.

[4] Y. Su and F. Jelinek, "Exploiting prosodic breaks in language modeling with random forests," in *Speech Prosody*, 2008.

[5] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Eurospeech*, 2005.

[6] M. Hoque, M. Sorower, M. Yeasin, and M. Louwerse, "What speech tells us about discourse: The role of prosodic and discourse features in dialogue act classification," in *IJCNN*, 2007.

[7] A. Rosenberg and J. Hirschberg, "Detecting pitch accents at the word, syllable and vowel level," in *HLT-NAACL*, 2009.

[8] M. Avanzi, N. Obin, A. Lacheret-Dujour, and B. Victorri, "Toward a continuous modeling of french prosodic structure: Using acoustic features to predict prominence location and prominence degree," in *Interspeech*, 2011.

[9] F. Tamburini and P. Wagner, "On automatic prominence detection for german," in *Interspeech*, 2007.

[10] F. Tamburini, "Automatic prominence identification and prosodic typology," in *Proc. InterSpeech 2005*, 2005, pp. 1813–1816.

[11] A. Maier, et al., "A language-independent feature set for the automatic evaluation of prosody," in *Interspeech*, 2009.

[12] A. Rosenberg, J. Hirschberg, and K. Manis, "Perception of english prominence by native mandarin chinese speakers," in *Speech Prosody*, 2010.

[13] J. Hirschberg and C. Nakatani, "A prosodic analysis of discourse segments in direction-giving monologues," in *Proc. of the 34th conference on Association for Computational Linguistics*, 1996, pp. 286–293.

[14] M. Avanzi,et al., "C-prom. an annotated corpus for french prominence studies," in *Proceedings of Speech Prosody 2010 Prosodic Prominence Workshop*, 2010.

[15] K. Eckart, A. Riester, and K. Schweitzer, "A discourse information radio news database for linguistic analysis," in *Linked Data in Linguistics*, 2012.

[16] A. Rosenberg, "Autobi – a tool for automatic tobi annotation," in *Interspeech*, 2010.

[17] ——, "Automatic detection and classification of prosodic events," Ph.D. dissertation, Columbia University, 2009.

[18] J. Hershey and P. Olsen, "Approximating the kullback-liebler divergence between gaussian mixture models," in *ICASSP*, 2007.

[19] H. DaumeIII, "Frustratingly easy domain adaptation," in *ACL*, 2007.