DISCOURSE CONNECTIVE DETECTION IN SPOKEN CONVERSATIONS

Giuseppe Riccardi, Evgeny A. Stepanov, Shammur Absar Chowdhury

Signals and Interactive Systems Lab
Department of Information Engineering and Computer Science
University of Trento, Trento, Italy

ABSTRACT

Discourse parsing is an important task in Language Understanding with applications to human-human and human-machine communication modeling. However, most of the research has focused on written text, and parsers heavily rely on syntactic parsers that themselves have low performance on dialog data. In our work, we address the problem of analyzing the semantic relations between discourse units in human-human spoken conversations. In particular, in this paper we focus on the detection of discourse connectives which are the predicate of such relations. The discourse relations are drawn from the Penn Discourse Treebank annotation model and adapted to a domain-specific Italian human-human spoken conversations. We study the relevance of lexical and acoustic context in predicting discourse connectives. We observe that both lexical and acoustic context have mixed effect on the prediction of specific connectives. While the oracle of using lexical and acoustic contextual feature combinations is $F_1 = 68.53$, the lexical context alone significantly outperforms the baseline by more than 10 points with $F_1 = 64.93$.

***Index Terms*—** Discourse Analysis, Speech Processing, Machine Learning

1. INTRODUCTION

Discourse parsing has application in many language technology areas that have to deal with units of data beyond sentence boundary. Such applications include Spoken Language Understanding (SLU), since turn may be composed of several sentences, and Spoken Dialog Systems (SDS), where dialog strategies are organized across several turns [1]. Mainly due to the lack of discourse annotated dialog data, most of the research on discourse parsing has focused on written text; and discourse parsers heavily rely on features extracted from syntactic parse trees (e.g. [2, 3, 4]). Unfortunately, syntactic parsers trained on written text behave poorly on dialog data [5], since the latter contain disfluencies and no sentence segmentation. In this paper we present exploratory experiments on discourse connective detection – initial step in Penn Discourse Treebank (PDTB) [6] style discourse parsing – in Italian spoken dialogs using acoustic and lexical features.

In the PDTB corpus, discourse relations are binary: the discourse connective and its two arguments *Arg1* and *Arg2*. *Arg2* is syntactically attached to the connective and *Arg1* is the other argument as shown in Example (1) in Figure 1, where *Arg2* is in **bold** and *Arg1* is in *italics*. Relations signaled by discourse connectives – members of the closed class – are *explicit* discourse relations. Detection of discourse connectives from English text using syntactic features has a very high performance ($F_1 = 94.19$) [7]. To our knowledge, this is the first work on the detection of discourse connectives from speech.

(1) [*In questo momento il palazzo non è collegato*]$_{Arg1}$
Allora [**è meglio collegarlo**]$_{Arg2}$
([*At this moment the building is not connected*]$_{Arg1}$
So [**we'd better connect it**]$_{Arg2}$)

(2) Allora vediamo un po' ecco qua
(So let's see here it is)

Fig. 1. Examples of *allora* (*so*) used as a discourse connective (1) and as a discourse marker (2).

Detection of discourse connectives from spoken dialogs is more challenging than from written text. In addition to the coordination of non-discourse units and polysemy, which occur both in dialogs and written text, in spontaneous conversations words that can function as discourse connectives can also function as *discourse markers* [8]. While discourse connectives relate discourse units, discourse markers are used for discourse organization and turn management, e.g. *allora* (so) in Example (1) is a discourse connective and in (2) it is a discourse marker [1]. Our goal is to discriminate between discourse connective category of a word token and all other usages.

In this paper we cast discourse connective detection as a binary classification task using lexical and acoustic features. We focus on the 10 most frequent Italian discourse connectives in the LUNA corpus [9] of human-human spoken conversations. Since our goal is to explore the relevance of acoustic and lexical context for the task, we experiment with features extracted from connective candidate spans and their left and right contexts in the window of ±2 tokens. We observe that both lexical and acoustic context have mixed effect on the detection of specific connectives.

The rest of the paper is structured as follows. In Section 2 we describe the data set used in the experiments, i.e. the Italian LUNA corpus. Then, in Section 3 we describe data pre-processing and feature extraction methodologies along with the features used for supervised machine learning. In Section 4 we describe classification experiments and obtained results. Section 5 provides concluding remarks and future directions.

2. DATA SET

The Italian LUNA Human-Human Corpus [9] is a collection of 572 dialogs in the hardware/software help desk domain. A subset of 60

The research leading to these results has received funding from the European Union – Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 610916 – SENSEI.

Connective		Data Freq.		ASR Freq.	
e	(and)	160	15.2%	154	96.2%
perchè	(because)	138	13.1%	136	98.6%
allora	(so)	91	8.7%	87	95.6%
però	(but)	87	8.3%	86	98.9%
ma	(but)	71	6.7%	71	100%
quindi	(then/so)	69	6.6%	67	97.1%
poi	(then)	62	5.9%	59	95.3%
se	(if)	60	5.7%	58	96.7%
così	(so)	33	3.1%	31	93.9%
che	(that)	25	2.4%	23	92.0%
Top 10		796	75.7%	772	96.3%
Rest (75)		256	24.3%		
Total (85)		1,052	100%		

Table 1. The 10 most frequent connectives in the LUNA Corpus and their % from total of *explicit* relations. ASR Freq. column gives % of connectives recognized by ASR.

Word	02: Train				03: Test			
	CONN		O		CONN		O	
e	97	43.3%	127	56.7%	40	46.0%	47	54.0%
perchè	81	83.5%	16	16.5%	35	83.3%	7	16.7%
allora	61	16.3%	313	83.7%	20	16.5%	101	83.5%
però	58	89.2%	7	10.8%	14	63.6%	8	36.4%
ma	45	56.3%	35	43.8%	16	55.2%	13	44.8%
quindi	44	57.9%	32	42.1%	17	51.5%	16	48.5%
poi	45	63.4%	26	36.6%	6	37.5%	10	62.5%
se	30	26.5%	83	73.5%	17	36.2%	30	63.8%
così	20	44.4%	25	55.6%	7	35.0%	13	65.0%
che	11	3.5%	302	96.5%	12	8.6%	127	91.4%
*	492	33.7%	966	66.3%	184	33.1%	372	66.9%

Table 2. Distribution of the 10 most frequent connectives (CONN) in training and test sets with the frequencies of their non-discourse connective usages (O).

dialogs was annotated [1] for discourse relations following Penn Discourse Treebank (PDTB) [6] guidelines. Out of total 1,606 annotated discourse relations, 1,052 are *explicit* discourse relations, that are signaled by 85 unique discourse connectives. In this paper we focus only on the 10 most frequent ones that are listed in Table 1 with their frequencies in data (Data Freq. column). This set of connectives accounts for 75.7% of all annotated explicit discourse relations. Some of the listed connectives additionally occur as tokens of other multi-word connectives (e.g. *che* is part of *visto che*). The amount of such multi-word connectives is 6.2%, most frequent being *che* (4.5%). To reduce noise, these multi-word connectives are removed from data.

The data (60 dialogs) is split into training, development and test splits as 42, 6, and 12 dialogs respectively. The distribution of the selected connectives into training and test sets after data pre-processing (described in the following Section) is given in Table 2.

3. FEATURE EXTRACTION

In this section we first describe data pre-processing; then feature extraction and the features themselves.

3.1. Data Pre-processing

The discourse annotation of the LUNA corpus was done using text extracted from manual transcriptions that do not contain word beginning and end time information. Thus, in order to be able to extract acoustic features for connective candidates, the text is aligned with the speech signal. The boundaries of words in the speech signal are obtained using forced alignment between word-level manual transcription and the speech signal within the manually segmented turn of a dialog. For the forced alignment, we use Automatic Speech Recognizer (ASR) that was trained on the LUNA Human-Human corpus using Kaldi [10] with Speaker Adaptive Training (SAT). The Word Error Rate (WER) of the ASR on the LUNA Human-Human test set is 39.7%. The ASR system is also used to produce automated transcriptions of manually segmented turns to match with discourse connectives in the corpus. Due to the error rate of the ASR, about 3.7% of discourse connectives are not recognized and are removed from data (see Table 1). After that, forced alignment is used to extract the features discussed in the following subsection.

3.2. Features

Several sets of features are extracted for supervised machine learning from the force-aligned ASR output. Lexical features are tokens and, since we train connective specific models, they appear only as lexical context features. The difference between manual and ASR output context tokens is negligible (lower that 0.1%). The rest of the features used in the experiments is described below.

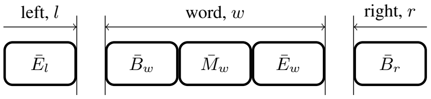
Duration and Silence Features: The time it took to utter a connective candidate and the duration of pauses before and after it might also carry information relevant to discourse. Thus, word and silence durations are extracted from the forced alignment and used as features for classification (3 features).

Acoustic Features: Acoustic frame-wise Low-Level Descriptors (LLD) are extracted using openSMILE [11] with the Frame Size = 25 ms and Frame Step = 10 ms. The extracted LLD are *prosodic* features (3) – fundamental frequency ($F0$), pitch, and loudness, all with their derivatives (2 per feature) – and *spectral* features (2) – flux and centroid (11 LLD in total).

We consider 3 segments – connective candidate token (w) and its left (l) and right (r) contexts (up to 2 words taken as a single segment), and acoustic features are extracted for each segment separately. In each segment, per frame feature values are normalized by Z-score with speaker-based mean ($\bar{m}_{spk}$) and its standard deviation (σ_{spk}), which are calculated using all the dialog turns of a corresponding speaker that do not contain discourse connective candidates. For normalization purposes overlapping turns are considered as a separate speaker.

Each segment S is later split into three parts – beginning (B_S), middle (M_S) and end (E_S); and arithmetic mean of all the frame feature values is calculated for the segment parts: $\bar{B}_S$, $\bar{M}_S$, and $\bar{E}_S$. As a result, there are 9 values per LLD for a connective candidate: means of beginning, middle and end parts for a candidate itself and its right and left contexts. Consequently, each candidate is represented by 99 acoustic features (11 LLD * 9 parts).

Acoustic Difference Features: In order to capture changes in the prosody within a word or with respect to context, using the acoustic features described above, we generate four acoustic difference features. For intra-word variation we calculate two differences – between middle ($\bar{M}_w$) and beginning ($\bar{B}_w$) and between end ($\bar{E}_w$) and middle ($\bar{M}_w$) parts of the word segment. For cross-word variation, the computed differences are between beginning part of the word ($\bar{B}_w$) and the final part of left context ($\bar{E}_l$) and between beginning



$$f_{i,1} = \bar{M}_w - \bar{B}_w;$$
$$f_{i,2} = \bar{E}_w - \bar{M}_w;$$
$$f_{i,3} = \begin{cases} \bar{B}_w - \bar{E}_l & : l \neq \emptyset \\ \bar{B}_w - \bar{m}_{spk} & : \text{otherwise} \end{cases} ;$$
$$f_{i,4} = \begin{cases} \bar{B}_r - \bar{E}_w & : r \neq \emptyset \\ \bar{m}_{spk} - \bar{E}_w & : \text{otherwise} \end{cases} ;$$

Fig. 2. Acoustic difference feature generation. Intra-word variation is represented by features $f_{i,1}$ and $f_{i,2}$ and cross-word variation by $f_{i,3}$ and $f_{i,4}$; where i is a Low-Level Descriptor (LLD).

part of the right context ($\bar{B}_r$) and the end part of the word ($\bar{E}_w$). In case of missing left or right context, the difference is computed with respect to the speaker mean $\bar{m}_{spk}$ (See Figure 2). The difference features are computed for each of 11 LLD; consequently, there are 44 difference features in total (11 LDD * 4 differences).

4. EXPERIMENTS AND RESULTS

Our goal is to study the relevance of the lexical and acoustic contexts for discourse connective detection from speech. The task is cast as binary discourse connective *vs.* all classification using acoustic and lexical features. Context is defined as features extracted from the segments to the left and right of a connective candidate (i.e. $S \in \{l, r\}$). Since pauses before and after word are not in the word segment, they are considered as context.

For classification we use AdaBoost algorithm [12] implemented in icsiboost [13]. All models are trained on 1,000 iterations, and, despite the unbalanced nature of our data, we do not apply any balancing techniques.

We describe four sets of experiments: (1) using acoustic features from only connective candidate segment (i.e. *without* context); (2) using acoustic features from only context segments (i.e. *from* context); (3) using acoustic features from all the segments (i.e. *with* context); and (4) using lexical context in isolation and with acoustic features. For settings 1-3 we train and evaluate models on the three sets of features described above – durations, acoustic features, and acoustic difference features – and their combination through vector fusion. For setting 4 we fuse lexical context with the fused vectors from settings 1-3.

Standard precision, recall and F_1 are used as evaluation metrics; however, due to space considerations, we report only F_1. We also compute a micro-averaged F_1 for whole connective set and test it for statistical significance. Statistical significance is measured using McNemar's χ^2 test with Yates' correction.

4.1. The Baseline

The baseline of discourse connective detection is computed as a majority decision. That is, if a word is more frequently appears as a connective in the training set, it is labeled as such in the test set. While some connectives have relatively high baselines (e.g. *perchè*: $F_1 = 90.91$ and *però*: $F_1 = 77.78$), the micro-averaged F_1 is low (53.99) since frequent discourse connectives *e* and *allora* mostly appear in non-discourse roles. For comparison, token only model on PDTB yields $F_1 = 75.33$ [7].

Conn.	**BL**	**D**	**Ac**$_N$	**Diff**$_N$	**ALL**$_N$
e	0.00	11.32	25.71	**46.58**	36.62
perchè	90.91	88.00	89.19	84.93	83.33
allora	0.00	0.00	5.71	**17.02**	11.76
però	77.78	77.78	74.29	74.29	68.75
ma	71.11	68.75	60.00	64.52	64.52
quindi	68.00	61.90	51.28	54.05	66.67
poi	54.55	47.06	52.63	46.15	33.33
se	0.00	21.43	37.50	25.00	**38.89**
così	0.00	28.57	**62.50**	40.00	50.00
che	0.00	0.00	0.00	0.00	0.00
Micro	53.99	50.46	49.86	51.37	51.54

Table 3. F_1 of models trained using non-contextual features: duration (**D**), acoustic features (**Ac**$_N$) and intra-word acoustic difference features (**Diff**$_N$) in isolation and in combination (**ALL**$_N$). **BL** is the majority baseline.

An alternative to training per connective models is a binary classification pooling all connectives together. The majority baseline for models trained using only connective tokens is identical for both settings. In preliminary pooled evaluation, only lexical context features have produced models outperforming the baseline. Thus, we focus on connective specific models and evaluate the relevance of acoustic and lexical context for each connective separately.

4.2. Connective Detection *without* Acoustic Context

Connective detection *without* context implies not using features outside of the connective candidate time frame; thus, they are word duration, acoustic features extracted from the word segment and within-word acoustic difference features, and their combinations. Results are reported in Table 3. All micro-averaged F_1, except for duration model (**D**), are significantly lower than the baseline. However, we observe that all the features contribute to the detection of a specific connective. Specifically, to the detection of the connectives mostly having non-discourse usages (i.e. *e*, *allora*, *se*, and *così*). Fusion of the features (**ALL**$_N$) does not produce the best model for all, but *se*.

4.3. Connective Detection *from* Acoustic Context

Connective detection *from* context implies using only features extracted from the left and right context of a connective candidate and their fusion. Context also includes lexical tokens in the window ±2, which are evaluated separately. Micro-averaged F_1 in Table 4, even thought often higher, are not significantly different from the baseline. Neither they are significantly different from the setting without context. Similar to the setting without context, the acoustic context features do contribute to the detection of the connectives mostly having non-discourse usages; however, they also contribute to the detection of others. Fusion of the features (**ALL**$_C$) does not produce the best model for all, but *così*.

4.4. Connective Detection *with* Acoustic Context

Connective detection *with* context implies that we can use all the features. Micro-averaged F_1 in Table 5 are not significantly different from either baseline or the previous settings, for all but acoustic

Conn.	**BL**	**S**	**Ac**$_C$	**Diff**$_C$	**ALL**$_C$
e	0.00	36.92	**56.47**	48.84	43.04
perchè	90.91	89.19	**91.89**	80.00	90.67
allora	0.00	0.00	15.38	**19.35**	13.79
però	77.78	77.78	**87.50**	74.29	77.78
ma	71.11	**71.43**	55.56	62.50	58.82
quindi	68.00	68.09	48.48	**70.97**	64.52
poi	54.55	54.55	47.06	15.38	37.50
se	0.00	0.00	**33.33**	**33.33**	22.22
così	0.00	0.00	42.11	53.33	**58.82**
che	0.00	0.00	0.00	**11.11**	0.00
Micro	53.99	55.49	**56.45**	53.41	55.06

Table 4. F_1 of models trained on only acoustic contextual features: silence durations (**S**), acoustic features from context (**Ac**$_C$), and cross-word acoustic difference (**Diff**$_C$) in isolation and in combination (**ALL**$_C$). **BL** is the majority baseline.

Conn.	**BL**	**DS**	**Ac**	**Diff**	**ALL**
e	0.00	33.85	**48.00**	36.62	32.43
perchè	90.91	84.93	89.47	84.06	90.67
allora	0.00	0.00	17.65	**18.75**	7.69
però	77.78	74.29	77.78	76.47	74.29
ma	71.11	76.47	68.57	**77.78**	62.50
quindi	68.00	61.11	60.61	64.52	**70.27**
poi	54.55	47.06	**66.67**	50.00	40.00
se	0.00	10.00	**42.86**	18.18	14.81
così	0.00	28.57	52.63	50.00	**53.33**
che	0.00	0.00	0.00	**13.33**	0.00
Micro	53.99	52.12	**58.95**	53.33	52.87

Table 5. F_1 of models trained on both contextual and non-contextual features: word and silence durations (**DS**), all acoustic features (**Ac**), all acoustic difference features (**Diff**), and their vector fusion (**ALL**). **BL** is the majority baseline.

Conn.	**BL**	**L**$_1$	**ALL**$_N$	**ALL**$_C$	**ALL**	***O***
e	0.00	**57.97**	39.47	39.02	37.84	*57.97*
perchè	90.91	**91.89**	87.67	87.67	90.67	*91.89*
allora	0.00	16.67	14.29	**31.25**	7.69	*31.25*
però	77.78	66.67	72.73	77.78	74.29	*87.70*
ma	71.11	**78.05**	55.17	50.00	66.67	*78.05*
quindi	68.00	44.44	66.67	66.67	**74.29**	*74.29*
poi	54.55	**57.14**	33.33	28.57	40.00	*66.67*
se	0.00	40.00	**47.06**	23.08	20.69	*47.06*
così	0.00	**66.67**	50.00	**66.67**	42.86	*66.67*
che	0.00	**86.96**	0.00	0.00	0.00	*86.96*
Micro	53.99	**64.93**	54.29	54.08	54.60	***68.53***

Table 6. F_1 of models trained on using lexical context in the window of ±1 tokens (**L**$_1$) in isolation and in combination with other features: acoustic features without context (**ALL**$_N$), acoustic features from context (**ALL**$_C$), and all acoustic features (**ALL**). **BL** is the majority baseline and **O** is the oracle of the best performing connective specific models.

features: F_1 for all acoustic features (**Ac**) is significantly higher than for the models without context (**Ac**$_C$). Similar to the previous settings, there are individual contributions to specific connectives and the fusion produces the best model only for *così*.

4.5. Connective Detection with Lexical Context

In this setting we evaluate the relevance of lexical context in isolation and through vector fusion with all speech derived features (durations, acoustic and acoustic difference) in the previous three settings. The lexical context is tokens in the window of ± 1 or ± 2 tokens. Results are reported in Table 6 (for space considerations we report only ± 1 window performance). The micro-averaged F_1 for lexical context in the window of ± 1 tokens performs significantly better than the baseline, while in the window of ± 2 is not. Thus, ± 1 window is used for vector fusion with other features.

The addition of lexical context to the speech-derived features does not produce significant changes to micro-averaged F_1. All acoustic-lexical models are not significantly different from the baseline or their equivalents without lexical context. The lexical context model with ± 1 token window is significantly better than the rest.

However, we again observe that individual connective performances are boosted. For connectives *allora*, *quindi* and *se* the fusion of acoustic features with lexical context produces the best results. In order to estimate the upper bound of the model combination (which could be achieved using features selection on the development set) we calculate the oracle of the best models per connective (**O**). The micro-averaged F_1 of the oracle is 68.53.

Overall, we observe that connectives behave differently with respect to acoustic and lexical context. Half of the connectives (*e*, *perchè*, *ma*, *che*, *così*) achieve the best results using only lexical context. The rest is quite diverse: *se* using lexical, but not acoustic context; *allora* using lexical and acoustic context without the features from the word segment, *quindi* using both lexical and acoustic contexts with the features from the word segment. The remaining two connectives *poi* and *però* perform better without lexical context: *però* using just acoustic features from context and *poi* using all the acoustic features; no acoustic difference or duration features for both.

All these differences lead us to conclude that discourse connectives are not uniform and different features are required to distinguish them from their non-discourse connective usages.

5. CONCLUSION

In this paper we have addressed the task of discourse connective detection from speech. We have focused on the relevance of lexical and acoustic context in discriminating the 10 most frequent discourse connectives from their non-discourse usages. We have observed that both lexical and acoustic context have mixed effect on the task. While lexical context model significantly outperforms the baseline with $F_1 = 64.93$, the oracle of combination with acoustic context has $F_1 = 68.53$. The conclusion is that the task of discourse connective detection is hard, but lexical features provide enough discriminative power to get improvement of more than 10 points over the majority baseline.

The future directions of this work are to investigate whether the reported observations are a property of Italian or spontaneous dialog. Given the oracle performance, data balancing and automatic features selection techniques may boost the overall performance.

6. REFERENCES

[1] Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind K. Joshi, "Annotation of discourse relations for conversational spoken dialogs.," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2010.

[2] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan, "A pdtb-styled end-to-end discourse parser," *Natural Language Engineering*, vol. 1, pp. 1 – 35, 2012.

[3] Evgeny A. Stepanov and Giuseppe Riccardi, "Comparative evaluation of argument extraction algorithms in discourse relation parsing," in *The 13th International Conference on Parsing Technologies (IWPT 2013)*, Nara, Japan, November 2013, pp. 36–44.

[4] Evgeny A. Stepanov, Giuseppe Riccardi, and Ali Orkan Bayer, "The UniTN discourse parser in CoNLL 2015 shared task: Token-level sequence labeling with argument-specific models," in *The SIGNLL Conference on Computational Natural Language Learning (CoNLL)- Shared Task*, Beijing, China, July 2015, pp. 25–31, ACL.

[5] Frederic Bechet, Alexis Nasr, and Benoit Favre, "Adapting dependency parsing to spontaneous speech for open domain spoken language understanding," in *Interspeech*, 2014.

[6] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber, "The penn discourse treebank 2.0," in *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[7] Emily Pitler and Ani Nenkova, "Using syntax to disambiguate explicit discourse connectives in text," in *Proceedings of the ACL-IJCNLP Conference*, 2009.

[8] Deborah Schiffrin, *Discourse Markers*, Cambridge University Press, 1987.

[9] Marco Dinarelli, Silvia Quarteroni, Sara Tonelli, Alessandro Moschitti, and Giuseppe Riccardi, "Annotating spoken dialogs: from speech segments to dialog acts and frame semantics," in *Proceedings of EACL Workshop on the Semantic Representation of Spoken Language*, Athens, Greece, 2009.

[10] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. 2011, IEEE.

[11] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, New York, NY, USA, 2010, pp. 1459–1462, ACM.

[12] Yoav Freund and Robert E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, August 1997.

[13] Benoit Favre, Dilek Hakkani-Tür, and Sebastien Cuendet, "Icsiboost," `https://github.com/benob/icsiboost/`, 2007.