A Context-Aware NLP Approach For Noteworthiness Detection in Cellphone Conversations

Francesca Bonin *
Trinity College Dublin
Dublin, Ireland
boninf@tcd.ie

Jose San Pedro
Telefonica Research
Barcelona, Spain
jspw@tid.es

Nuria Oliver
Telefonica Research
Barcelona, Spain
nuriao@tid.es

Abstract

This papers presents a context-aware NLP approach to automatically detect noteworthy information in spontaneous mobile phone conversations. The proposed method uses a supervised modeling strategy which considers both features from the content of the conversation as well as contextual information from the call. We empirically analyze the predictive performance of features of different nature on a corpus of mobile phone conversations. The results of this study reveal that the context of the conversation plays a crucial role on boosting the predictive performance of the model.

1 Introduction

More than 6 billion people worldwide use their cellphones daily for a variety of purposes: contacting colleagues, relatives or friends, doing business, getting help in emergency situations, etc. Previous work (Carrascal et al., 2012) has shown that almost 40% of users frequently feel the need to recall bits of information from their phone conversations and that 27% of the users consider the recall task to be difficult, mainly because taking notes during a mobile phone call is not always possible (*e.g.* hands not free, lack of time or devices for note-taking). In a related user study, Cycyl *et al.* reveal that users are often engaged in concurrent tasks during mobile phone conversations (*e.g.* walking, jogging, driving, cooking, etc), which makes taking notes an unfeasible task (Cycil et al., 2013).

In this setting, information extraction techniques could be applied to automatically detect noteworthy information from mobile phone conversations. Related studies have focused on detecting noteworthiness from meeting transcripts (Banerjee and Rudnicky, 2009). However, very little work has been done to date to identify this kind of information in other types of human communication, such as spontaneous phone conversations.

In this paper, we present a data-driven information extraction approach aimed at automatically detecting fragments of phone conversations worth annotating for future recall, *i.e. noteworthy*. These *call notes* could then be presented to the users to enable fast browsing of their conversation history, and leveraged to design efficient information interaction techniques for supporting smart user interfaces.

Given the particular characteristics of mobile phone calls, detecting noteworthiness in them is challenging at many levels. First, the audio is captured in a natural environment rather than in controlled settings, which results in noisy signals, and consequently in noisy transcriptions. Second, the conversations are highly fragmented due to their spontaneous nature. Finally, at a conceptual level, judging which pieces of information are noteworthy is a very subjective task, as emerged in (Banerjee and Rudnicky, 2009), who investigated the feasibility of the task by conducting a Wizard of Oz-based user study.

Our noteworthiness modeling approach considers a supervised learning paradigm which takes into account two types of information: (1) Contextual information both from the call (*where, when, to whom,* ...) and the users (*gender, age,* ...); and (2) Content information of the conversation. The combination

* The work was conducted while the author was intern at Telefonica Research, Barcelona, Spain.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 25–36, Dublin, Ireland, August 23-29 2014.

of both sources of information enhances the flexibility of the model to accurately predict noteworthiness in different use scenarios.

The main contributions of this paper are:

i) We propose and evaluate a supervised machine learning model to automatically detect noteworthy segments of phone conversations. Our approach adopts a hybrid strategy to model conversations exploiting both *content* and *context*-related information.

ii) We propose a new set of content and context-based features specifically designed to detect noteworthy information in our corpus of real-world cellphone conversations, and compare their effectiveness

iii) We provide a discussion of the results, derived from our quantitative and qualitative analyses.

The paper is structured as follows. Relevant previous work is presented in Section 2. Section 3 describes the corpus of phone conversations and the annotations provided by the participants. In Section 4 we describe in-depth the extracted features. Our experimental validation and results are presented in Section 5. Finally, Section 6 summarizes our findings and highlights some lines of future research.

2 Related work

Noteworthiness detection in conversations can be considered to be a particular form of summarization: the aim is to summarize the conversation by keeping only the *relevant* pieces of information that the user would like to refer to at a later time. Although related, the main distinction between automatic summarization and detection of noteworthy information lays in the notion of *relevance*. The former aims at generating a comprehensive record of the conversation, while the latter considers only fragments worth registering for future recall.

Considerable research activity has recently been devoted to automatic text and speech summarization (Maskey and Hirschberg, 2003). Many approaches have been proposed in the literature, including cluster (Zhang et al., 2005) and graph-based methods (Garg et al., 2009; Wang and Liu, 2011) and machine learning techniques (Jian Zhang et al., 2007; Maskey and Hirschberg, 2006; Galley, 2006), where the task is tackled as a binary classification problem considering whether the sentence is a good candidate for a summary or not. In addition, different types of features have been used, including lexical, acoustic and structural characteristics (Xie et al., 2008; Maskey and Hirschberg, 2005). Recent works have been focused on adapting summarization to the social context, exploiting user generated contents associated with the documents (Yang et al., 2011; Hu et al., 2012). Implicit and explicit community feedback in online collaborative websites have also been leveraged to detect highlights of media assets (San Pedro et al., 2009).

However, few studies have focused on noteworthiness detection. Banerjee *et al.* investigate the feasibility of discovering noteworthy pieces of information in meetings by means of a Wizard of Oz-based user study where a human suggested notes to meeting participants during the meeting. The authors found that the human annotator obtained a precision of 35% and a recall of 41.5%. In the same work, Banerjee *et al.* reports a low inter annotator agreement (IAA) in noteworthiness discovery. In a related work –probably the most relevant prior-art to our work, the authors apply extractive meeting summarization techniques to automatically detect noteworthy utterances in meetings (Banerjee and Rudnicky, 2008). They train a Decision Tree classifier over a collection of 5 meetings, obtaining an F-score of 0.14. This result highlights the difficulty of the task at hand and motivates to explore alternative approaches.

To overcome the difficulties posed by this task we propose two main contributions: 1) the use of novel features engineered ad-hoc for this task, and 2) the use of contextual information. While the former adapts the document representation to the specific problem setting, the latter allows to enhance the representation with orthogonal information which many times provides a higher discriminative power. This approach has been used successfully in related fields; for instance, in information retrieval tasks rich multimodal queries have been shown to effectively boost the retrieval performance compared to pure textual queries (Yeh et al., 2011).

3 Corpus Collection

We used a corpus of cellphone conversations collected in a previous study (Carrascal et al., 2012). In this study, a large sample of mobile phone conversations was recorded, semi-automatically transcribed[1] and manually annotated for relevance by their participants. Over 64 days, 796 mobile phone conversations from 62 volunteering subjects (20 female) were recorded. All the participants were Spanish native speakers, and the conversations were recorded and transcribed in Spanish. Metadata about the call (*e.g.* duration, date, time) was also stored along with the actual conversation and its transcript. More details about the corpus collection process can be found in (Carrascal et al., 2012).

All the participants were first asked to fill out a pre-study questionnaire where they provided some personal information, including gender, marital status, education and income. Then they were asked to annotate what parts of their calls that they would like to take a note of: *i.e.* noteworthy fragments of conversations. To this end, participants used a Web-based interface that gave them access to their calls and allowed them to highlight with the mouse the parts of the transcript that they considered to be worth keeping for future reference.

We used these annotations as the ground truth for the studies presented in this paper, considering them as the ideal noteworthy parts of the calls. For privacy reasons, due to the sensible nature of the data (i.e. private phone conversations) we could not consider alternative ground truth generation schemes, for instance collecting annotations from users other than the callers themselves.[2]

Finally, the participants were asked to fill out a questionnaire after annotating each call, which was used to collect contextual information, including: location of the call (*i.e at work, at home, while commuting, while doing shopping, while exercising*), and category of the call (*i.e. discuss a topic, taking an appointment, give/receive information, asking a favor, social*).

3.1 Characteristics of the Corpus

The original conversation collection consists of a total of 796 conversations, of an average length of 178 seconds ($s = 384$ sec.). We pre-filtered this original set to exclude calls with problems in the transcript (*e.g.* empty transcript, only one speaker audible, etc). Out of the entire corpus we finally selected 659 conversations. We denote this subset of the corpus as the $\mathcal{G}$ dataset. The $\mathcal{G}$ dataset comprises $22,474$ turns, with an average of 34.10 ($s = 45$) turns per conversation. From these, only 671 are annotated as being noteworthy (2.98%), which represent an average of 1.02 turns ($s = 1.803$) per call. Given that the vast majority of turns (97.2%) are not annotated, this can be considered a highly unbalanced dataset, which makes the automatic modeling problem more challenging.

Hence, we considered a second dataset which included only the 295 calls from the $\mathcal{G}$ dataset containing at least one annotation. This second subset, denoted as $\mathcal{A}$ amounts for approximately 45% of the $\mathcal{G}$ dataset. The $\mathcal{A}$ dataset features $10,642$ turns, with an average of 36.07 ($s = 33$) turns per conversation. From these, again 671 (6.3%) are annotated, which represent an average of 2.275 ($s = 2.09$) per call. The $\mathcal{A}$ dataset is still highly unbalanced but significantly less than the $\mathcal{G}$ dataset. Table 1 summarizes the high level characteristics of each dataset.

		Turns		Annotated Turns	
	# Calls	Total	$avg.$ per call	Total	Fraction
$\mathcal{G}$	659	$22,474$	34.1 ($s = 45$)	671	2.9%
$\mathcal{A}$	295	$10,642$	36 ($s = 33$)	671	6.3%

Table 1: General statistics on $\mathcal{G}$ and $\mathcal{A}$ datasets.

Class	Annotations
I	*We are in front of the fruit shop*
RoA	*Tomorrow we go to look for the swimsuit*
RI	*Are you coming to eat? At what time*
O	*Sure, it's normal*

Table 2: Examples of annotations.

Given the complexity of the modeling problem, we studied the note taking behavior of participants to identify relevant patterns that would simplify the problem. To this end, we conducted a quantitative analysis of the note taking behavior of participants. We found that users tend to highlight complete

[1] Participants were given the opportunity to revise transcriptions during the annotation phase.

[2] Receivers of the calls were aware of the study and were given the possibility to not participate in the call, but were not directed involved in the study.

turns as relevant, instead of parts of the turns. On average, 66.57% ($s = 35.87$) of the words within an annotated turn are highlighted, with a median value of 80%. Hence, we decided to use *turns* –rather than individual words– as the unit to be automatically detected as noteworthy. Using this approach, a turn is considered to be noteworthy if it contains at least one annotated word.

3.2 Qualitative Analysis of the Corpus

Since our aim is to detect the noteworthy turns within a call, we conducted a preliminary qualitative analysis to understand the nature of the annotations entered by the participants in the study. We distinguished 4 types of annotations: *Giving Information (I)*, *Requesting Information (RI)*, *Reporting on an Action (RoA)* and *Other (O)*. Examples of these 4 types of annotations are presented in Table 2. We collected annotations from three collaborators of our lab for a total of 54 randomly selected turns from the $\mathcal{A}$ dataset (IAA, $Fleiss\ Kappa = 0.54$ (Fleiss, 1971)).

We found that 47% of the turns were classified as belonging to the *Giving Information* category, 22% of the turns to the *Request Information* category, 26% to the *Other* category, and only 3% were classified as *Report on an Action*. Intuitively, we had expected the *Giving Information* category to be the most common in the annotated turns. However, the results obtained show that the other types of annotations are also well represented in the data.

Two main interesting aspects emerge. First, while the vast majority of annotations correspond to turns where a piece of information is given (*e.g. We meet at 3pm*), turns where information is requested are also well represented in the sample. There are plausible explanations for this behavior, such as users trying to include more context in the annotations. Second, more than 25% of this manually annotated dataset was marked under the *Other* category, which includes turns with very diverse functionalities (*e.g.* greetings, statements of agreement). This reveals that participants tend to annotate turns with very diverse functional aspects, which poses a challenge to be added to the unbalanced nature of the dataset.

4 Feature Extraction

We follow a supervised machine learning approach to automatically detect noteworthy turns in conversations. In this section we describe the features that we compute to represent conversations and which have been engineered to capture information relevant to the problem at hand. We have divided the set of features into two categories: **C**ontent features, that we denote with the letter **C**, and conte**X**t features, that we denote with the letter **X**.

4.1 Content Features

Content features are computed by analyzing the content of the conversations. We use as input the textual information resulting from the semi-automatic transcription of the calls. Note that we do not make use of any conversational acoustic information. While the analysis of the acoustic signal may reveal additional cues useful for noteworthiness detection, it lies out of the scope of this work.

In order to extract features from the transcript, we first pre-proces the datasets (split in turns, lemmatized, PoS tagged). Also, we extract and classify Named Entities (NEs).[3] We extract 42 content-based features which include both variations of features previously used in the meeting summarization literature and novel features particularly adapted to our task. However, in contrast to related work on meeting summarization, we do not extract content features based on lexical similarity to the entire call or to the main topic of the call, under the intuition that the notion of noteworthiness depends on the user's needs rather than on the main topic of the conversation. In addition and for robusteness purposes, we decided not to rely on long distance dependency information (*e.g.* argument predicate relations) or deep syntactical parsing, which are sensitive to the quality of the transcription.

The resulting features are grouped into three main classes: **Turn-Based** (C-T), **Dynamic** (C-D), and **Conversational** (C-C). We compare them with a pure bag-of-words (BoW) representation. Table 3a provides a summary of all the content-based features used in our system. Where applicable, we experi-

[3] All pre-processing was performed using the Freeling Language Processing tools (Padro et al., 2010).

ment with two vector representations: binary and frequency-based. We will refer to these two different encoding schemes as **Bin** for the binary case, and **Freq** for the frequency case.

CONTENT FEATURES	
C-BoW (Bag of Words)	
BoW	BoW for all words (except hapax)
C-T (Turn-based)	
NE	Presence (or frequency) of NEs (Person, Location, Organization, Numbers, Dates, Misc.)
TLN	Turn length in # words normalized
PoS	PoS distribution
TF	Max and Mean term frequency
IDF	Max and Mean inverse document frequency
C-D (Dynamic)	
Rep	Repetition between t and t-1,t+1,t-2,t+2
Int	Presence (or total amount) of Int. pro./adj. in t-1
Q	Presence (or total amount) of question in t-1
C-C (Conversational)	
Dur	Duration of the call (# turns and # words)
Cent	Conversation centrality
Spk	Speaker
Dom	Speaker dominance

(a) Content Feature

CONTEXT FEATURES	
X-C (Call-based)	
X-C-T	Time of the call
X-C-Loc	Location of the call
X-C-Day	Day of the call
X-C-Obj	Objective of the call
X-U (User-based)	
X-U-G	Gender
X-U-A	Age
X-U -I	Income
X-U-E	Education
X-U-Ms	Marital Status

(b) Context Feature

Table 3: Content (a) and Context (b) based features.

4.1.1 Turn-Based Content features (C-T)

Turn-based content features take into account information related to individual turns. We distinguish lexical and non-lexical C-T.

Lexical content features: Lexical C-T features capture the lexical properties of a turn. We include NEs, such as *Locations, Organizations, Persons*, *Miscs* and *Numbers*, *Dates*, and temporal expressions. For each turn t, we detect the presence of any NE as well as the presence of individual classes of NEs. For each of these class of entities, we extract both a binary and a frequency feature vector. In the text summarization literature, the appearance of particular lexical phrases (*e.g. to summarize*) has been exploited to predict relevant sentences (Gupta and Lehal, 2010). In our study, attention has been given to the presence of temporal expressions under the intuition that temporal cues are good indicators of upcoming pieces of information (*e.g. The meeting is tomorrow*). We exploit temporal expressions, such as *today, tomorrow, etc*.[4]

Non-lexical content features: capture characteristics of the turn which do not involve lexical information, namely: turn length, Part-of-Speech (PoS) distributions and Tf-Idf descriptive statistics at the turn level.

In meeting summarization, the average length of a turn has been found to be a good feature to automatically create a summary of a meeting (Xie et al., 2008). In our dataset, preliminary analyses revealed that annotated turns tend to be longer in average. Hence, we include the turn length in the non-lexical content feature set. The turn length is given by the number of tokens per turn normalized over the average turn

[4]Note that, here and in the remainder of the paper, we report the English translations of the Spanish originals.

length (punctuation excluded). To further gauge discourse characteristics, we detect the distribution of PoS at the turn level: *i.e.* for each turn, the frequency of nouns, pronouns, adjectives, adverbs, interjections, verbs, prepositions and conjunctions is calculated. Finally, we compute the term frequency (Tf) and inverse document frequency (Idf) measures. In (Xie et al., 2008), authors report that Idf is among the most discriminative features in sentence selection for text summarization. We compute maximum and mean Tf and Idf values for each turn.

4.1.2 Dynamic content features (C-D)

Dynamic content features are designed to capture the semantic relationships between each turn and its precedent and subsequent turns. In particular we refer to relations such as lexical and topical cohesion, question-answer relationship, and the appearance of general cues that may anticipate relevant bits of information in the subsequent turn. We consider: *1)* the lexical and topical cohesion among consecutive turns (Repetitions); *2)* the appearance of general cues that may anticipate relevant bits of information in the subsequent turn (Interrogative Pronouns); *3)* the question-answer relationship among consecutive turns (Question).

Repetitions: words repeated by different speakers in consecutive turns. Participants of a conversation tend to align at several linguistic and paralinguistic levels in order to ease communication and increase mutual understanding (Pickering and Ferreira, 2008). This phenomenon has been investigated in terms of prosody, lexicon and syntax (Levitan and Hirschberg, 2011; Brennan, 1996; Bonin et al., 2013; Branigan et al., 2010). From a lexical point of view, the alignment mechanism, often referred to as priming, is realized by means of word repetitions among speakers. Many studies have investigated this phenomenon assessing correlation between priming and mutual understanding or dialogue success (Vogel, 2013; Reitter and Moore, 2007).

We exploit the priming phenomenon to detect concepts in the conversation that are considered important by both participants, relying on the fact that repeated words convey concepts that participants want to make sure they have been successfully communicated to their interlocutor. Given a dataset D, a turn in D, $t \in D$, and $t - i$ and $t + i$ turns in the context of t, we calculate the amount of repeated lemmas between t and $t-i$, and t and $t+i$ for $1 \leq i \leq 2$. In order to consider semantically meaningful repetitions, we take into account only content words (nouns, adjectives, adverbs, verbs) when they activate one of the C-T features described above. Being A the set of annotated turns, we noticed a significant difference in the amount of repeated lemmas between t,$t - i$ for $t \in A$ rather than for $t \notin A$. Find below an example of consecutive turns with repetitions:

```
Turn Utterance
  t-1: Starting at half past four.
  t:   Starting at half past four, yes.
```

Interrogative pronouns and questions: We also exploit indicators of an upcoming *giving information* act. As shown in Sec 3.2, 47% of the annotations were marked as *giving information*, which may have been triggered by a request of information in the precedent turn. Hence, in order to capture these cases, we identify linguistic elements that indicate a request of information in $t - 1$ (questions and interrogative pronouns/adjectives).

4.1.3 Conversational flow features (C-C)

They are designed to model information about the conversation's flow and speakers' interaction.

Centrality of the turn: Distance of a turn from the center of the conversation. This feature is inspired by the sentence location features used in text summarization (Chen et al., 2002). Chen *et al.* assign different weights to sentences in the first, middle and final part of a paragraph, in order to favor sentences that are in the central part of the paragraph as they are considered to be more informative for a summary. In our corpus, we noticed the tendency of users to annotate turns that are in the central portion of the conversation. Typically the first and the last quarters of the phone conversations are dedicated to social talk. Hence, we introduce a temporal feature, referred to as conversation centrality, that captures the distance of a turn from the center of the conversation. This distance is measured in terms of number of words, excluding punctuation.

Speaker: Who is uttering the turn (caller vs callee).

Conversation duration: Length of the conversation in number of turns and in number of words. The number of turns captures the dynamics of a dialogue (few longer turn vs a more dynamic exchange), while the number of words captures the overall duration.

Speaker dominance: We consider whether the speaker is the dominant speaker of the conversation, defining dominance in terms of amount of productions during the call. This is calculated by comparing the number of turns of speaker *a* vs speaker *b*, normalized over the total amount of turns per call.

4.1.4 Bag-of-Words (BoW)

Finally, we explore the performance of a naive bag-of-words scheme to represent the content at the turn level. Given the large vocabulary size of our corpus ($10,144$ tokens) and the sparsity organic to bag-of-word representations, we decided to use a trivial dimensionality reduction strategy filtering out the terms that appear only once in the corpus. We decided not to apply a stop-list of functional words for further reducing the feature space. This decision was based on the higher discrimintative power we observed when comparing classification accuracy with and without them. We discarded the use of more aggressive feature selection approaches (*e.g.* mutual information) to allow for a fair comparison of accuracy with the rest of feature representations described in the paper. In total, our BoW representation had $5,048$ dimensions in the $\mathcal{G}$ dataset, and $3,219$ dimensions in the $\mathcal{A}$ dataset.

4.2 Context Features

Context features are introduced under the assumption that noteworthy information may depend on the characteristics of the user and on the situation in which the call takes place. For example, people may not need to annotate pieces of information that are part of their daily lives. Whereas while taking an appointment, it is plausible the need to annotate the name of the doctor, in a social call with a friend, the name of the friend is part of the background knowledge of the user. Therefore, while from a content (and an NLP) point of view both names are Person NEs and carry the same amount of information, from the point of view of the user they might have different weight (no need of taking note vs need of taking note). Also, the current situation or location of the user may influence the necessity of taking notes: a user in a supermarket will not need to annotate to buy milk, (s)he will rather take it directly from the shelf. A user driving to the supermarket will need to keep in his/her mind the need to buy milk for later recall.

In line with this, we noted in Section 2 that pure NLP approaches applied to automatically detecting noteworthy information in meetings are able to achieve an F-score of only 0.14. This low F-score underlines the complexity of the task and the limitations of a pure content-based approach. Contextual cues may be used to increase the discriminative power of the classification model.

Since we consider the specific scenario of cellphone conversations, we can exploit contextual information derived from the use of the mobile network, such as geo-location, and temporal information. Other contextual features that we use, gathered during the pre-study questionaire, are organically much more challenging to infer. We still decided to consider these as a way to assess the potential of several types of contextual information with respect to the discriminative power of the classifier. We distinguish among Call-based (X-C) and User-based (X-U) contextual features. A schematic overview of these features is given in Table 3b.

4.2.1 Call-Based Features (X-C)

Call-based features are meant to capture contextual information at the call level. In particular, X-C features include information about *where*, *when* and *for what reason* a call is made, under the intuition that calls made, for example, during working hours may have different noteworthy information than calls made in the weekend. We distinguish six *location* categories: home, work place, while commuting, while exercising, while shopping, other. The location of the calls was provided by participants through the post-call questionnaire. However location information is typically available from the mobile network. In terms of *temporal* features, we consider the actual time of the call (over 24 hours). In addition, we classify the time in two classes: working vs non working hours, and the day in also two classes: weekday vs weekend. Finally, we also consider the *objective* of the call as described in Section 3. Note that,

although this information is not directly accessible from mobile data collected during the call, previous literature on conversation classification supports the feasibility of inferring this information from the content of the conversation (Koço et al., 2012).

4.2.2 User-Based Features (X-U)

Finally, we introduce a set of features that feed the model with information about the user. We exploit information that could be provided by users upon registration to such a *note-taking* service. We capture age, gender, educational level, income and marital status. Gender is represented as a binary feature, while age is categorized in 5 groups: below 20 years old, between 20 and 30, between 30 and 40, between 40 and 50 and above 50. The education status is represented by the following categories: Primary education, Secondary education, Bachelor degree or a Postgraduate education (Master or PhD). Yearly income is categorized by: up to 10k, 20k, 30k, 40k and more than 40k. Finally, marital status is categorized as: single, in a couple (married, with a stable partner), other.

5 Experiments

The goal of our system is to automatically identify information annotated by users in terms of its potential need for future recall. We frame this problem as a binary classification task (noteworthy or not) at the turn level. This task presents two main challenges. First, our dataset is extremely unbalanced, with less than 3% of the corpus labeled as relevant by the participants. Second, the subjectivity of the task leads to high variability of annotation behaviours, (see Sec. 3.2). In this section we describe the experimental setting that we used to empirically evaluate the performance of different features sets and present the results obtained using the ground truth data collected (Section 3) to provide classification performance scores. In order to fully investigate the predictive performance of the different feature sets, we conducted our experiments using both the entire corpus $\mathcal{G}$, which includes all the selected conversations, and its subset $\mathcal{A}$, which considers only the calls with at least one annotation. Both sets are described in Sec. 3. We experimented using both encoding schemes described in Section 4: binary based (**Bin**) and frequency based (**Freq**).

We used Support Vector Machines (SVMs) with RBF kernel, as this classification approach yielded the most consistent results throughout all the evaluated configurations. We used the same random split of training and test sets for all the experiments, accounting for 70% and 30% of the dataset respectively. We tune the hyperparameter C of the SVM model using a 3-fold cross-validation approach on the training data only, where we chose F-score as the quality metric to optimize. Given the nature of our task, recall is preferred to precision from a user-centric perspective: it is preferable to avoid missing any relevant information than to include some non-relevant fragments. For this reason, we also report precision and recall values.

5.1 Classification Results

This section presents the results obtained in our binary classification task (turns being noteworthy or not). We study the performance of different combinations of features and present the results obtained using only content information (C), and the combination of content and context information (CX).

5.1.1 Content features

We present a comparison of the different content feature sets using the naming scheme of Sec.4. We considered four classification scenarios: C-T only, C-D only, the combination of C-T and C-D (C-TD), and the combination of C-T, C-D and C-C (C-TDC). The results of these feature sets are shown in Tables 4a and 4b for the $\mathcal{G}$ and $\mathcal{A}$ collections, respectively.

As shown in Table 4a, the maximum F-score for the $\mathcal{G}$ dataset is achieved for the combination of all content features included the BoW. The low score ($F = 0.18$) is a direct consequence of the low precision obtained ($P = 0.11$). For the $\mathcal{A}$ dataset (Table 4b) we observe a better F-score ($F = 0.296$), still obtained by the combination of all content features, with a much higher precision ($P = 0.18$) due to the significant amount of noise removed by considering only annotated calls. Note that in both the $\mathcal{G}$ and the $\mathcal{A}$ datasets the C-TDC feature set outperforms the pure BoW approach ($F = 0.158$ vs. $F = 0.14$

Features	Precision		Recall		F-score	
Encoding	Bin	Freq	Bin	Freq	Bin	Freq
BoW	0.081	0.083	0.730	0.720	0.150	0.150
C-T	0.087	0.088	0.53	0.32	0.15	0.139
C-D	0.03	0.26	0.26	0.12	0.15	0.05
C-TD	0.087	0.09	**0.754**	0.33	0.1505	0.1419
C-TDC	0.09	0.093	0.58	0.37	0.158	0.149
C-TDC+BoW	**0.11**	**0.11**	0.52	0.51	**0.18**	**0.18**

(a) Results for the $\mathcal{G}$ dataset.

Features	Precision		Recall		F-score	
Encoding	Bin	Freq	Bin	Freq	Bin	Freq
BoW	0.14	0.14	0.54	0.54	0.22	0.22
C-T	0.135	0.147	0.56	0.555	0.218	0.23
C-D	0.149	0.11	0.24	0.15	0.18	0.12
C-TD	0.143	0.143	0.506	0.546	0.223	0.227
C-TDC	0.165	0.159	0.626	**0.693**	0.254	0.267
C-TDC+BoW	**0.188**	0.1659	0.57	0.568	0.283	**0.296**

(b) Results for the $\mathcal{A}$ dataset.

Table 4: Classification performance of Content features, BoW and their combination.

for $\mathcal{G}$ and $F = 0.267$ vs. $F = 0.22$ for $\mathcal{A}$), using a fraction (about 1%) of the number of BoW features, which leads to a considerably simpler model. On the other hand, the combination of C-TDC and BoW features improves the results up to $F = 0.18$ for $\mathcal{G}$ ($P = 0.11$, $R = 0.52$) and $F = 0.296$ ($P = 0.20$, $R = 0.57$) for the $\mathcal{A}$ subset. This result highlights how the lexical representation comprised by the BoW provides the model with orthogonal information to the one provided by the C-TDC features set.

To the best of our knowledge no previous work has been done in noteworthy detection from telephone conversations. For this reason, we report as a reference the results of the more similar prior art to our work, (Banerjee and Rudnicky, 2008), where the authors implement an SVM classifier for the detection of noteworthy information in meetings.[5] Although aware of the different nature of the dataset, these results are reported to get a sense of the potentiality of the system. The best performance of our model on the $\mathcal{A}$ dataset improves in 15% the F-score of $F = 0.14$ reported in (Banerjee and Rudnicky, 2008).

5.1.2 Combining Content and Context Features

In this section we report the performance of the model trained using both content and context features. For simplicity, in the remainder of this section we refer to the entire set of content features, (C-TDC) as C, to the entire set of context features as X, and to their combination as CX. When we test adding BoW features the *+BoW* naming is used. The results are shown in Table 5 and Figure 1. We observe that the fusion of content and context features (CX and CX+BoW) provides a noticeable overall increase in the F-score for both datasets. This increase is particularly high for the $\mathcal{G}$ dataset, where the F-score gets increased by almost a factor of 2, from $F = 0.18$ to $F = 0.28$. On the $\mathcal{A}$ dataset, the combination of content and context features improves the F-score from $F = 0.29$ to $F = 0.32$, given by a better precision ($P = 0.24$ vs $P = 0.18$) with similar recall.

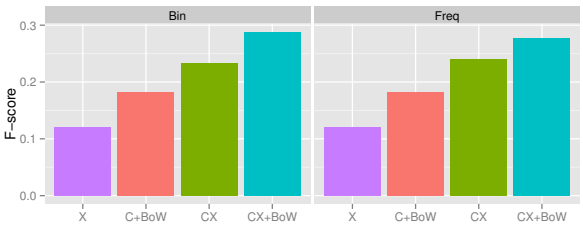
Features	Precision		Recall		F-score	
Rep.	B	F	B	F	B	F
C+BoW	0.11	0.11	0.52	0.51	0.18	0.18
X	0.068	0.068	**0.665**	**0.665**	0.124	0.124
CX	0.169	**0.20**	0.38	0.286	0.2354	0.2394
CX+BoW	0.189	0.1919	0.524	0.5022	**0.288**	0.277

(a) Results for the $\mathcal{G}$ dataset.

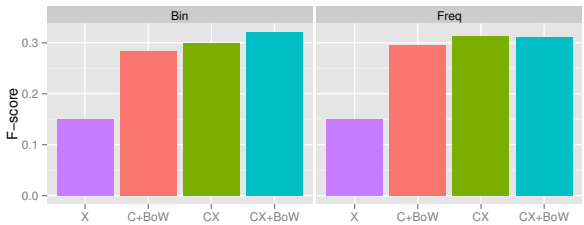
Features	Precision		Recall		F-score	
Rep.	B	F	B	F	B	F
C+BoW	0.188	0.1659	0.57	0.568	0.283	0.296
X	0.087	0.087	0.56	0.56	0.15	0.15
CX	0.2075	0.212	0.5866	**0.595**	0.3066	0.3130
CX+BoW	0.223	**0.2455**	0.573	0.426	**0.3212**	0.3116

(b) Results for the $\mathcal{A}$ dataset.

Table 5: Classification performance using the combination of context and context-based features



(a) Results for the $\mathcal{G}$ dataset.



(b) Results for the $\mathcal{A}$ dataset.

Figure 1: Classification performance using Content, Context features and their combination

This result gives empirical evidence that these two sets of features convey complementary information

[5] In their experimental settings all the meetings have at least one annotation as in our $\mathcal{A}$ scenario.

that is relevant for the task at hand. That is, the same words can carry different relevance depending on the contextual information of the conversation.

Note that the BoW features add discriminative information in the $\mathcal{G}$ scenario, but have a minimal effect in the less noisy $\mathcal{A}$ scenario where the combination of content and context features, without BoW, provides already an F-score of $F = 0.31$.

An interesting remark about this combined model is that the difference in performance between the $\mathcal{G}$ and the $\mathcal{A}$ dataset is vastly reduced. While in the pure content model the difference in F-score value between both datasets was 0.10, in the combined model this difference is just 0.03. This result shows that the combination of content and context features boosts considerably the results in the more noisy and realistic dataset $\mathcal{G}$, while its effect is weaker in the cleaner dataset.

5.2 Qualitative Analysis

In order to better understand the failure cases in our system, we carried out a qualitative analysis of both false positives, *i.e.* turns annotated by the system but not by the user, and false negatives, *i.e.* turns annotated by the user but not by the system. Table 6 illustrates a few representative examples. Note how the proposed system does not perform well when detecting a *request for information* as something worth annotating (*e.g. What are you doing?*). We noticed that in these cases, the model tended to annotate turns where the information was actually provided (*e.g.That is the package has arrived*). We can hypothesize that users annotate the *request for information* to give context to the a-priori more relevant information, *i.e.* the answer to the question. However, in some cases, participants did not annotate the answer as relevant. This counter-intuitive observation reflects the subjectivity and variability of the task.

False Positive	False Negative
I am leaving soon, I start at 3 o'clock or [..]	*How are you? Can you hear me?*
Let's see if we can tell him.	*What are you doing?*
That is, the package has arrived.	*Did you buy beautiful things for me?*

Table 6: Examples of false positive and false negative turns.

5.3 Comparative Analysis and Discussion

To the best of our knowledge there are no previous works of similar nature to the study presented in this paper. Yet, it is important to give a sense of the merits and limitations of the proposed approach in the context of the state-of-the-art. For this reason, we compare our results with (Banerjee and Rudnicky, 2009), which is the most similar prior art to our work. In (Banerjee and Rudnicky, 2009), Barnejee *et al.* perform a Wizard-of-Oz experiment and report a performance of the human annotator of $P = 0.35$ precision, $R = 0.42$ recall, leading to an F-score of $F = 0.38$. This result highlights the difficulty of the task, even for a human annotator. When comparing our proposed system with this Wizard-of-Oz experiment, we obtain an F-score of $F = 0.32$ against the human annotator's F-score of $F = 0.38$, with a significantly higher recall (0.57 vs 0.42) yet lower precision (0.245 vs 0.35). Given this human-based prediction performance, the proposed approach represents a good first step towards realizing an intelligent annotation system for mobile phone conversations.

6 Conclusions and Future Work

In this paper we have proposed and empirically evaluated a machine learning-based approach to automatically detect noteworthy information in spontaneous mobile phone conversations. The subjectivity of this task leads to a challenging classification problem even for human assessors. Our approach adopts a hybrid strategy that exploits the content and the context of the conversation. We have shown that information about the context of the conversation improves the predictive performance of the system over a pure content based approach.

In the future, we plan to extend the model by including acoustic features which could improve the performance by adding orthogonal information to the current model. To tackle the subjectivity of the task we also intend to investigate the performance of personalization techniques, creating individual models per user. Finally, we plan to conduct a study to evaluate our system from a user-centric perspective.

Acknowledgments

This work was partially supported by the Innovation Bursary of Trinity College Dublin (project: 'Technology for Harmonising Interpersonal Communication').

References

Satanjeev Banerjee and Alexander I. Rudnicky. 2008. An extractive-summarization baseline for the automatic detection of noteworthy utterances in multi-party human-human dialog. In *Proceeding of SLT*, pages 177–180.

Satanjeev Banerjee and Alexander Rudnicky. 2009. Detecting the noteworthiness of utterances in human meetings. In *Proceedings of the SIGDIAL 2009 Conference*, pages 71–78, London, UK, September. Association for Computational Linguistics.

Francesca Bonin, Celine De Looze, Sucheta Ghosh, Emer Gilmartin, Carl Vogel, Anna Polychroniou, Hugues Salamin, Alessandro Vinciarelli, and Nick Campbell. 2013. Investigating fine temporal dynamics of prosodic and lexical accommodation. In *Proceedings of Interspeech 2013*, Lyon, France, August.

H.P. Branigan, M.J. Pickering, J. Pearson, and J.F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.

Susan E. Brennan. 1996. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, pages 41–44.

Juan Pablo Carrascal, Rodrigo de Oliveira, and Mauro Cherubini. 2012. A note paper on note-taking: understanding annotations of mobile phone calls. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, MobileHCI '12, pages 21–24, New York, NY, USA. ACM.

Fang Chen, Kesong Han, and Guilin Chen. 2002. An approach to sentence-selection-based text summarization. In *TENCON'02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering*, volume 1, pages 489–493. IEEE.

Chandrika Cycil, Mark Perry, Eric Laurier, and Alex Taylor. 2013. 'eyes free' in-car assistance: parent and child passenger collaboration during phone calls. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, MobileHCI '13, pages 332–341, New York, NY, USA. ACM.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Michel Galley. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 364–372, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nikhil Garg, Benoit Favre, and Dilek Hakkani-Tür. 2009. Clusterrank: a graph based method for meeting summarization. In *Technical Report, IDIAP*.

Vishal Gupta and Gurpreet Singh Lehal. 2010. A survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2(3):258–268.

Po Hu, Dong-Hong Ji, Chong Teng, and Yujing Guo. 2012. Context-enhanced personalized social summarization. In *COLING*, pages 1223–1238.

J Jian Zhang, Ho Yin Chan, and Pascale Fung. 2007. Improving lecture speech summarization using rhetorical information. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 195–200. IEEE.

Sokol Koço, Cécile Capponi, and Frédéric Béchet. 2012. Applying multiview learning algorithms to human-human conversation classification. In *INTERSPEECH*.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Twelfth Annual Conference of the International Speech Communication Association*.

Sameer Maskey and Julia Hirschberg. 2003. Automatic summarization of broadcast news using structural features. In *INTERSPEECH*.

Sameer Maskey and Julia Hirschberg. 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *INTERSPEECH*, pages 621–624.

Sameer Maskey and Julia Hirschberg. 2006. Summarizing speech without text using hidden markov models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 89–92. Association for Computational Linguistics.

Lluis Padro, Samuel Reese, Eneko Agirre, and Aitor Soroa. 2010. Semantic services in freeling 2.1: Wordnet and ukb. In Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen, editors, *Principles, Construction, and Application of Multilingual Wordnets*, pages 99–105, Mumbai, India, February. Global Wordnet Conference 2010, Narosa Publishing House.

Martin J. Pickering and Victor S. Ferreira. 2008. Structural priming: a critical review. *Psychological bulletin*, 134(3):427.

David Reitter and Johanna D Moore. 2007. Predicting success in dialogue. In *Annual Meeting-Association for Computational Linguistics*, volume 45, page 808.

Jose San Pedro, Vaiva Kalnikaite, and Steve Whittaker. 2009. You can play that again: Exploring social redundancy to derive highlight regions in videos. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, IUI '09, pages 469–474, New York, NY, USA. ACM.

Carl Vogel. 2013. Attribution of mutual understanding. *Journal of Law & Policy*, pages 101–145.

Dong Wang and Yang Liu. 2011. A pilot study of opinion summarization in conversations. In *Annual Meeting-Association for Computational Linguistics, ACL*, pages 331–339.

Shasha Xie, Yang Liu, and Hui Lin. 2008. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pages 157–160. IEEE.

Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. 2011. Social context summarization. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 255–264, New York, NY, USA. ACM.

Tom Yeh, Brandyn White, Jose San Pedro, Boriz Katz, and Larry S. Davis. 2011. A case for query by image and text content: Searching computer help using screenshots and keywords. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 775–784, New York, NY, USA. ACM.

Yongzheng Zhang, Nur Zincir-Heywood, and Evangelos Milios. 2005. Narrative text classification for automatic key phrase extraction in web document corpora. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, WIDM '05, pages 51–58, New York, NY, USA. ACM.