Hello, Who is Calling?: Can Words Reveal the Social Nature of Conversations?

Anthony Stark, Izhak Shafran and Jeffrey Kaye
Center for Spoken Language Understanding, OHSU, Portland USA.
{starkan,shafrani,kaye}@ohsu.edu

Abstract

This study aims to infer the social nature of conversations from their content automatically. To place this work in context, our motivation stems from the need to understand how social disengagement affects cognitive decline or depression among older adults. For this purpose, we collected a comprehensive and naturalistic corpus comprising of all the incoming and outgoing telephone calls from 10 subjects over the duration of a year. As a first step, we learned a binary classifier to filter out business related conversation, achieving an accuracy of about 85%. This classification task provides a convenient tool to probe the nature of telephone conversations. We evaluated the utility of openings and closing in differentiating personal calls, and find that empirical results on a large corpus do not support the hypotheses by Schegloff and Sacks that personal conversations are marked by unique closing structures. For classifying different types of social relationships such as family vs other, we investigated features related to language use (entropy), hand-crafted dictionary (LIWC) and topics learned using unsupervised latent Dirichlet models (LDA). Our results show that the posteriors over topics from LDA provide consistently higher accuracy (60-81%) compared to LIWC or language use features in distinguishing different types of conversations.

1 Introduction

In recent years, there has been a growing interest in analyzing text in informal interactions such as in Internet chat, newsgroups and twitter. The emphasis of most such research has been in estimating network structure (Kwak et al., 2010) and detecting trending topics (Ritter et al., 2010), sentiments (Pak and Paroubek, 2010) and first stories (Petrović et al., 2010). The focus has been on aggregating information from large number of users to analyze population level statistics.

The study reported in this paper, in contrast, focuses on understanding the social interactions of an individual over long periods of time. Our motivation stems from the need to understand the factors of social engagement that ameliorate the rate of cognitive decline and depression in older adults. Since the early work of Glass (1997) and colleagues, several studies on large cohorts over extended duration have confirmed that older adults with few social relationships are at an increased risk of suffering depression and dementia. The limited information available in currently used coarse measures, often based on self-reports, have hindered epidemiologists from probing the nature of this association further.

While social engagement is typically multi-faceted, older adults, who are often less mobile, rely on telephone conversations to maintain their social relationships. This is reflected in a recent survey by Pew Research Center which reported that among adults 65 years and older, nine in ten talk with family or friends every day and more than 95% use land-line telephones for all or most of their calls (Taylor et al., June 29 2009). Conveniently for us, telephone conversations present several advantages for analysis. Unlike many other forms of communication, the interaction is restricted solely to an audio

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 112–119,
Montréal, Canada, June 3-8, 2012.

channel, without recourse to gestures or facial expressions. While we do not discount the importance of multi-modal communication, having a communication channel restricted to a unimodal format does significantly simplify both collection and analysis. Furthermore, the use of a handset affords the opportunity to capture naturalistic speech samples at relatively high signal-to-noise ratio. Lastly, automatic speech recognition (ASR) systems can now transcribe telephone conversations with sufficient accuracy for useful automated analysis.

Given the above premise, we focus our attention on studying social interactions of older adults over land-line telephones. To facilitate such a study, we collected telephone conversations from several older adults for approximately one year. Note that our corpus is unlike the publicly available Switchboard and Fisher corpora, which contain conversations between unfamiliar speakers discussing a topic from a pre-determined list such as music, crime, air pollution (Godfrey et al., 1992). In contrast, the conversations in our corpus are completely natural, covering a wide range of topics, conversational partners and types of interactions. Our corpus is also comprehensive in that it includes all the outgoing/incoming calls from subjects' homes during the observation period.

As a step toward understanding social networks and associated relationships, our first task was to classify social and non-social (business) conversations. While reverse listing was useful to a certain extent, we were unable to find listing on up to 50% of the calls in our corpus due to lack of caller ID information on many calls as well as unlisted numbers. Moreover, we cannot preclude the possibility that a social conversation may occur on a business number (e.g., a friend or a relative working in a business establishment) and vice versa. Using the subset of calls for which we have reliable listing, we learned a supervised classifier and then employed the classifier to label the remaining calls for further analysis.

The focus of this study was not so much on learning a binary classifier, but using the resulting classifier as a tool to probe the nature of telephone conversations as well as to test whether the scores obtained from it can serve as a proxy for degree of social familiarity. The classifier also affords us an opportunity to re-examine hypotheses proposed by Schegloff and Sacks (1974; 1968; 1973) about the structure of openings and closing in business and personal conversations. Within social conversation, we investigated the accuracy of identifying conversations with close friends and relatives from others.

The rest of this paper is arranged as follows. After describing the corpus and ASR system in Sections 2 and 3, we probe the nature of telephone conversations in Section 4. We present direct binary classification experiments in Section 5 and lastly, we close with a few remarks in Section 6.

2 Corpus: Everyday Telephone Conversations Spanning a Year

Our corpus consists of 12,067 digitized land-line telephone conversations. Recordings were taken from 10 volunteers, 79 years or older, over a period of approximately 12 months. Subjects were all native English speakers recruited from the USA. In addition to the conversations, our corpus includes a rich set of meta-data, such as call direction (incoming vs outgoing), time of call, duration and DTMF/caller ID when available. At the end of the data collection, for each subject, twenty telephone numbers were identified corresponding to top ten most frequent calls and top ten longest calls. Subjects were asked to identify their relationship with the speakers at these numbers as immediate family, near relatives, close friends, casual friends, strangers and business.

For this initial study, we discard conversations with less than 30 automatically transcribed words. This was done primarily to get rid of spurious and/or noisy recordings related to device failure as well as incorrectly dialed telephone numbers. Moreover, short conversations are less likely to provide enough social context to be useful.

Of the 8,558 available conversations, 2,728 were identified as residential conversations and 1,095 were identified as business conversations using reverse listings from multiple sources; e.g. phone directory lookup, exit interviews, internet lookup. This left 4,395 unlabeled records, for which the reverse listing was either inconclusive or for which the phone number information was missing and/or improperly recorded.

3 Automatic Speech Recognition

Conversations in our corpus were automatically transcribed using an ASR system. Our ASR system is structured after IBM's conversation telephony system which gave the top performance in the most recent evaluation of speech recognition technology for telephony by National Institute of Standards and Technology (Soltau et al., 2005). The acoustic models were trained on about 2000 hours of telephone speech from Switchboard and Fisher corpora (Godfrey et al., 1992). The system has a vocabulary of 47K and uses a trigram language model with about 10M n-grams, estimated from a mix of transcripts and web-harvested data. Decoding is performed in three stages using speaker-independent models, vocal-tract normalized models and speaker-adapted models. The three sets of models are similar in complexity with 4000 clustered pentaphone states and 150K Gaussians with diagonal covariances. Our system does not include discriminative training and performs at a word error rate of about 24% on NIST RT Dev04 which is comparable to state of the art performance for such systems. The privacy requirements in place for our corpus prohibit human listening – precluding the transcriptions needed reporting recognition accuracy. However, while our corpus differs from Switchboard, we expect the performance of the 2000 hour recognizer to be relatively close to results on NIST benchmark.

4 Nature of Telephone Conversations

4.1 Classification Experiments

As mentioned earlier, we first learned a baseline binary classifier to filter out business calls from residential calls. Apart from using this as a tool to probe the characteristics of social calls, it also helps us to classify unlabeled calls and thus avoid discard half the corpus from subsequent analysis of social network and relationships. Recall, the labels for the calls were obtained using reverse lookup from multiple sources. We assume that the majority of our training set reflect the true nature of the conversations and expect to employ the classifier subsequently for correcting the errors arising when personal conversations occur on business lines and vice versa.

We learned a baseline SVM classifier using a balanced training set. From the labeled records we created a balanced verification set containing 164,115 words over 328 conversations. The remainder was used to create a balanced training set consisting of 866,696 words over 1,862 conversations. The SVM was trained on 20-fold cross validation and evaluated on the verification set. After experimenting with different kernels, we found an RBF kernel to be most effective, achieving an accuracy of 87.50% on the verification data.

4.2 Can the Scores of the Binary Classifier Differentiate Types of Social Relationship?

Since the SVM score has utility in measuring a conversation on the social-business axis, we now examine its usefulness in differentiating social ties. To test this, we computed SVM score statistics for all conversations with family and friends. For comparison, we also computed the statistics for all conversations automatically tagged as residential as well as all conversations in the data. Table 1 shows the average family score is unambiguously higher than the average residential conversation (independent sample t-test, $p < 0.001$). This is an interesting result since distinction of family conversations (from general social calls) never factored into the SVM. Rather, it appears to arise naturally as an extrapolation from the more general residential/business discriminator. The friend sub-population exhibited statistics much closer to the general residential population and its differences were not significant to any degree. The overlap between scores for conversations with family and friends overlap significantly. Notably, the conversations with family have a significantly higher mean and a tighter variance than with other social ties.

Table 1: SVM scores for phone number sub-categories.

Category	# Calls	Mean score	STD
Family	1162	1.12	0.50
Friends	532	0.95	0.51
Residential	2728	0.93	0.63
Business	1095	-1.16	0.70
Global	8558	0.46	0.96

4.3 How Informative are Openings and Closings in Differentiating Telephone Conversations?

Schegloff and Sacks assert openings (beginnings) and closings (ends) of telephone conversations have certain identifiable structures (Sacks et al., 1974). For example, the structure of openings facilitate establishing identity of the conversants and the purpose of their call (Schegloff, 1968). Closings in personal conversations are likely to include a pre-closing signal that allows either party to mention any unmentioned mentionables before conversation ends (Schegloff and Sacks, 1973).

Given the above assertions, we expect openings and closings to be informative about the type of conversations. Using our classifier, we compare the accuracy of predicting the type from openings, closings and random segments of the conversations. For different lengths of the three types of segments, the observed performance of the classifier is plotted in Figure 1. The results for the random segment were computed by averaging over 100 trials. Several important results are immediately apparent. Openings possess much higher utility than closings. This is consistent with general intuition that the opening exchange is expected to clarify the nature and topic of the call. Closings were found to be only as informative as random segments from the conversations. This is contrary to what one might expect from Schegloff and Sack's assertion that pre-closing differ significantly in personal telephone calls (Schegloff and Sacks, 1973). Less intuitive is the fact that increasing the length of the opening segment does not improve performance. Surprisingly, a 30-word segment from the opening appears to be sufficient to achieve high classification accuracy (87.20%).

4.4 Data Sparsity or Inherent Ambiguity: Why are Short Conversations difficult to Classify?

Sparsity often has a deleterious effect on classification performance. In our experiments, we noticed that shorter conversations suffer from poor classification. However, the results from the above section appear to contradict this assertion, as a 30-word window can give very good performance. This seems to suggest short conversations suffer poor recognition

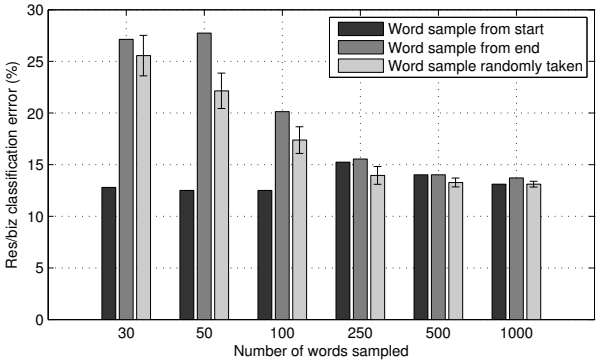

Figure 1: Comparison of classification accuracy in predicting the type of conversation from openings, closings and random segments. Error bars are one standard deviation.

due to properties beyond the obvious sparsity effect. To test this, we investigated the differences in short and long conversations in greater detail. We separate calls into quintile groups based on word counts. However, we now calculate all features from a 30-word opening – eliminating effects directly related to size. The results in Table 2 show that the ability to predict the type of conversation does not degrade when long conversations are truncated. Meanwhile, the accuracy of classification drops for (originally) short conversations. There is a surprisingly small performance loss due to the artificial truncation. These observations suggest that the long and short conversations are inherently different in nature, at least in their openings.

Table 2: Accuracy in predicting the type of conversation when they are truncated to 30-words of openings based on conversation length quintiles. The column, Res / Biz, split gives the label distributions for the quintiles.

Orig. Word Counts		Split	Accuracy
Quintile	#Words	Res. / Biz.	
0-20	30-87	62.12 / 37.88	78.6
20-40	88-167	48.48 / 51.52	82.8
40-60	168-295	39.39 / 60.61	91.4
60-80	296-740	40.91 / 59.09	87.8
80-100	741+	59.38 / 40.62	93.4

We should point out that spurious recordings in our corpus are concentrated in the low word count group – undoubtedly dropping their accuracies. However, the trend of improving accuracy persists well into the high word count ranges where spu-

rious records are rare. Given this fact, it appears that individuals in our corpus are more careful in enunciating the reasons for calling if an extended phone conversation is anticipated.

4.5 Can Openings Help Predict Relative Lengths of Conversations?

From the results presented so far, we know that openings are good predictors of the type of conversations yet to unfold. We also know that there are inherent language differences between short and long conversations. So, it is natural to ask whether openings can predict relative lengths of conversations. To test this hypothesis, we bin conversations into 5 groups or ranks based on their percentile lengths (word counts) – very short, short, moderate, long and very long durations, as in Table 2. Using independent features from the 30-word opening, we attempt to predict the relative rank of two conversations by learning a rank SVM (Joachims, 2006). We found the ranker to give 27% error rate, significantly lower (independent sample t-test, d.f. $\approx$ 1M, $p<0.01$) than the random chance of 40%. Chance baseline was determined using Monte Carlo simulation (1M random rankings) in conjunction with the rank SVM evaluation (Joachims, 2006).

Features from very short conversations may contain both openings and closings, i.e., both a *hello* and a *goodbye*, making them easier to rank. To avoid this confounding factor, we also compute performance after discarding the shortest grouping of conversations ($<$ 88 words) to ensure closings are avoided in the 30-word window. The resulting classifier over *short, medium, long, very long* conversations ranked 30% of the pairs erroneously, somewhat better than chance at 37%. Though the performance gain over the random ranker has shrunk considerably, there is still some utility in using the opening of a conversation to determine its ultimate duration. However, it is clear predicting duration via conversation opening is a much more difficult task overall.

5 Supervised Classification of Types of Social Relationships

While the scores of the binary classifier provided statistically significant differences between calls to different types of social relationships, they are not particularly useful in classifying the calls with high accuracy. In this section, we investigate the performance of classifiers to differentiate the following binary classes.

- *Residential vs business*
- *Family vs all other*
- *Family vs other residential*
- *Familiar vs non-familiar*

Familiar denotes calls to those numbers with whom subject has conversed more than 5 times. Recall that the numbers corresponding to family members were identified by the subjects in a post-collection interview. We learned binary classifier for the four cases, a few of which were reported in our early work (Stark et al., 2011). We investigated a variety of features in these tasks. A breakdown of the corpus is give in Table 3. Not all categories are mutually exclusive. For example the majority of *family* conversations also fall into the *familiar* and *residential* categories.

Table 3: Number of conversations per category.

Category	Instances
Biz.	1095
Residential	2728
Family	1111
Res. non-family	1462
Familiar	3010
All	8558

5.1 Lexical Statistics

Speakers who share close social ties are likely to engage in conversations on a wide variety of topics and this is likely to reflect in the entropy of their language use. We capture this aspect of language use by computing language entropy over the unigram word distribution for each conversation, i.e; $H(d) = -\sum_w p(w|d) \log p(w|d)$, where $p(w|d)$ is the probability of word w given conversation d. We also included two other lexical statistics namely the speaking rate and the word count (in log domain). Table 4 lists the utility of these language properties for differentiating the four binary classes mentioned earlier, where the p-value is computed using two tailed independent sample t-test.

Table 4: T-statistics for different context groups. Labels: a) Log-word count, b) speaking rate, c) language entropy. Asterisk denotes significance at p<0.0001. Sample sizes (n) may be found in Table 3.

Task	d.f.	a)	b)	c)
Res. v. biz.	7646	1.9	10.1^*	-1.9
Family v. other	8556	16.3^*	9.0^*	13.4^*
Family v. other res.	2571	12.9^*	5.1^*	11.3^*
Familiar v. other	8556	10.4^*	6.4^*	9.3^*

For the most part, the significance tests conform with preconceived ideas of language use over the telephone. It is shown that people talk longer, more rapidly and have wider range of language use when conversing with a familiar contact and/or family member. Surprisingly, only the speaking rate showed significant differences among the residential/business categories, with business conversations being conducted at a slower pace at least for the elderly demographic in our corpus.

5.2 Linguistic inquiry and Word Count

We investigated a hand-crafted dictionary of salient words, called *Linguistic Inquiry and Word Count* (LIWC), employed in social psychology studies (Pennebaker et al., 2003). This dictionary group words into 64 categories such as pronouns, activity words, positive emotion and health. The categories have significant overlap and a given word can map to zero or more categories. The clear benefit of LIWC is that the word categories have very clear and pre-labeled meanings. They suffer from the obvious drawback that the words are labeled in isolation without taking their context into account. The tags are not chosen under any mathematical criteria and so there are no guarantees the resultant feature will be useful or optimal for classifying utterances.

Table 5 lists the LIWC categories significant ($p< 0.001$) to the different classes. The listed terms are sorted according to their t-statistic, with early and later terms more indicative of first and second class labels respectively.

5.3 Latent Dirichlet allocation

Unsupervised clustering and feature selection can make use of data for which we have no labels. For example, in the case of business and residential labels, unlabeled data amounts to about 50% of our corpus. Motivated by this consideration, we examined unsupervised clustering using Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

LDA models a conversation as a bag of words. The model generates a conversation by: (a) sampling a topic distribution θ for the conversation using a per-conversation Dirichlet topic distribution with a hyper-parameter α, (b) sampling a topic z for each word in the conversation using a multinomial distribution using the topic mixture θ, and (c) sampling the word from a per-topic multinomial word distribution with a hyper-parameter β (Blei et al., 2003). The number of topics are assumed to be given. The per-conversation topic distribution and the per-topic word distribution can be automatically estimated to maximize the likelihood of training data. The sparsity of these two distributions can be controlled by tweaking α and β; lower values increase sparsity.

For our experiments, we estimated a maximum likelihood 30-topic LDA model from the corpus. Experimentally, we found best cross-validation results were obtained when α and β were set to 0.01 and 0.1 respectively.

When peering into the topics learned by the LDA method, it did appear that topics were approximately separated into contextual categories. Most interesting, when the number of clusters are reduced to two, the LDA model managed to segment residential and business conversations with relatively high accuracy (80%). This suggests the LDA model was able to approximately learn these classes in an unsupervised manner.

Table 6 lists words strongly associated with the two topics and clearly the unsupervised clustering appears to have automatically differentiated the business-oriented calls from the rest. On closer examination, we found that most of the probability was distributed in a limited number of words in the business-oriented topic. On the contrary, the probability was more widely distributed among words in the other cluster, reflecting the diversity of content in personal calls.

5.4 Classifying Types of Social Relationships

Though t-tests are useful for ruling out insignificant relationships, they are insufficient for quantifying the degree of separability – and thus, ultimately their

Table 5: LIWC categories found to be significant in classifying relationships, ranked according to their t-statistic.

Relationship	Categories
Res. v. biz.	I, Past, Self, Motion, Other, Insight, Eating, Pronoun, Down, Physcal, Excl, Space, Cogmech, Home, Sleep, Tentat, Assent, / Article, Optim, Fillers, Senses, Hear, We, Feel, Inhib, Incl, You, School, Money, Occup, Job, Number
Family v. all	Other, Past, Assent, Sleep, Insight, I, Pronoun, Cogmech, Tentat, Motion, Self / Affect, Optim, Certain, Future, School, Comm, Job, We, Preps, Incl, Occup, You, Number
Family v. res.	Other, Past, Sleep, Pronoun, Tentat, Cogmech, Insight, Humans / Comm, We, Incl, You, Preps, Number
Familiar v. other	Other, Assent, Past, I, Leisure, Self, Insight / Fillers, Certain, Social, Posemo, We, Future, Affect, Incl, Comm, Achieve, School, You, Optim, Job, Occup

Table 6: Per-topic word distribution learned using unsupervised clustering with LDA. Words are sorted according to their posterior topic distribution. Words with identical distributions are sorted alphabetically.

Topic 1	Topic 2
Invalid, helpline, eligibility, transactions, promotional, representative, mastercard, touchtone, activation, nominating, receiver, voicemail, digit, representatives, Chrysler, ballots, staggering, refills, resented, classics, metro, represented, administer, transfers, reselling, recommendations, explanation, floral, exclusive, submit.	Adorable, aeroplanes, Arlene, Astoria, baked, biscuits, bitches, blisters, bluegrass, bracelet, brains, bushes, calorie, casinos, Charlene, cheeses, chit, Chris, clam, clientele, cock, cookie, copying, crab, Davenport, debating, dementia, dictionary, dime, Disneyland, eek, Eileen, fascinated, follies, fry, gained.

utility in discrimination. To directly test discrimination performance, we use support vector machine classifiers. Before performing classification, we produce balanced datasets that have equal numbers of conversations for each category. Our primary motivation for artificially balancing the label distribution in each experiment is to provide a consistent baseline over which each classifier may be compared. We learn SVM classifiers with an RBF kernel using 85% of data for development. SVM parameters are tuned with 20-fold cross-validation on the devset. The accuracies of the classifiers, measured on a held out set, are reported in Table 7.

We tested four feature vectors: 1) unigram frequencies, 2) surface language features (log word count, speaking rate, entropy), 3) the 64 dimension LIWC frequency vector and 4) a 30-dimension vector of LDA topic posterior log-probabilities.

Table 7: SVM performance for the language features. Labels: a) unigram vector, b) lexical statistics, c) LIWC and d) LDA topic posterior log-probabilities

Task	1-grams	L.Stats	LIWC	LDA
Res. v. biz.	84.95	67.61	78.70	81.03
Family v. all	78.03	61.16	72.77	74.75
Family v. res.	76.13	62.92	71.06	72.37
Familiarity	69.17	60.92	64.20	69.56

Overall, the plain unigram frequency vector provided the best discrimination performance. However, this comes at significant training costs as the unigram feature vector has a dimensionality of approximately 20,000. While the surface features did possess a degree of classification utility, there are clearly outclassed by the content-based features. Furthermore, their integration into the content-features yielded only insignificant improvements to accuracy. Finally, it is of interest to note that the 30-topic LDA feature trained with ML criterion outperformed the 64-topic LIWC vector in all cases.

6 Conclusions

This paper studies a unique corpus of conversational telephone speech, a comprehensive and naturalistic sample of all the incoming and outgoing telephone calls from 10 older adults over the duration of one year. Through empirical experiments we show that the business calls can be separated from social calls with accuracies as high as 85% using standard techniques. Subgroups such as family can also be differentiated automatically with accuracies above 74%. When compared to language use (entropy) and hand-crafted dictionaries (LIWC), poste-

riors over topics computed using a latent Dirichlet model provide superior performance.

For the elderly demographic, openings of conversations were found to be more informative in classifying conversation than closings or random segments, when using automated transcripts. The high accuracy in classifying business from personal conversations suggests potential applications in designing context user interface for smartphones to offer icons related to work email, work calendar or Facebook apps. In future work, we plan to examine subject specific language use, turn taking and affect to further improve the classification of social calls (Shafran et al., 2003).

7 Acknowledgements

This research was supported in part by NIH Grants 5K25AG033723-02 and P30 AG024978-05 and NSF Awards 1027834, 0958585 and 0905095. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the NIH or NSF. We thank Brian Kingsbury and IBM for making their ASR software tools available to us. We are also grateful to Nicole Larimer, Maider Lehr and Katherine Wild for their contributions to data collection.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

T A Glass, C F Mendes de Leon, T E Seeman, and L F Berkman. 1997. Beyond single indicators of social networks: a lisrel analysis of social ties among the elderly. *Soc Sci Med*, 44(10):1503–1517.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520.

T. Joachims. 2006. Training linear svms in linear time. In *ACM Conference on Knowledge Discovery and Data Mining*.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 172–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation language. *Language*, 50(4(1)):696–735.

Emanuel A. Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica*, 8:289–327.

Emanuel A. Schegloff. 1968. Sequencing in conversational openings. *American Anthropologist*, 70(6):1075–1095.

Izhak Shafran, Michael Riley, and Mehryar Mohri. 2003. Voice signatures. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*.

H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig. 2005. The IBM 2004 conversational telephony system for rich transcription. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 205–208.

Anthony Stark, Izhak Shafran, and Jeffrey Kaye. 2011. Supervised and unsupervised feature selection for inferring social nature of telephone conversations from their content. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*.

Paul Taylor, Rich Morin, Kim Parker, D'Vera Cohn, and Wendy Wang. June 29, 2009. Growing old in America: Expectations vs. reality. http://pewsocialtrends.org/files/2010/10/Getting-Old-in-America.pdf.