Overview of the Evalita 2018 itaLIan Speech acT labEliNg (iLISTEN) Task

Pierpaolo Basile and **Nicole Novielli**
Università degli Studi di Bari Aldo Moro
Dipartimento di Informatica
Via E. Orabona, 4 - 70125 Bari (ITALY)
{pierpaolo.basile|nicole.novielli}@uniba.it

Abstract

English. We describe the first edition of the " itaLIan Speech acT labEliNg" (iLISTEN) task at the EVALITA 2018 campaign (Caselli et al., 2018). The task consists in automatically annotating dialogue turns with *speech act labels*, i.e. with the communicative intention of the speaker, such as statement, request for information, agreement, opinion expression, or general answer. The task is justified by the large number of applications that could benefit from automatic speech act annotation of natural language interactions such as tools for the intelligent information access, that is by relying on natural dialogues. We received two runs from two teams, one from academia and the other one from industry. In spite of the inherent complexity of the tasks, both systems largely outperformed the baseline.

Italiano. *Descriviamo la prima edizione del task di "itaLIan Speech acT labEliNg" (iLISTEN) organizzato nell'ambito della campagna di valutazione EVALITA 2018. Il task consiste nell'annotazione automatica di turni di dialogo con la label di speech act corrispondente. Ciascuna categoria di speech act denota l'intenzione comunicativa del parlante, ossia l'intenzione di formulare un'affermazione oggettiva, l'espressione di un'opinione, la richiesta di informazioni, una risposta, un'espressione di consenso. Riteniamo che il task sia rilevante per la il dominio della linguistica computazionale e non solo, alla luce del recente interesse da parte della comunità scientifica nei confronti dei paradigmi di interazione e accesso intelligente all'informazione basati su dialogo. Il task ha visto la partecipazione di due team, uno accademico e uno industriale. Nonostante la complessità del task proposto, entrabi i team hanno ampiamente superato la baseline.*

1 Introduction

Speech acts have been extensively investigated in linguistics (Austin, 1962; Searle, 1969), and computational linguistics (Traum, 2000; Stolcke et al., 2000) since long. Specifically, the task of automatic speech act recognition has been addressed leveraging both supervised (Stolcke et al., 2000; Vosoughi and Roy, 2016) and unsupervised approaches (Novielli and Strapparava, 2011). This interest is justified by the large number of applications that could benefit from automatic speech act annotation of natural language interactions.

In particular, a recent research trend has emerged to investigate methodologies to enable intelligent access to information, that is by relying on natural dialogues as interaction metaphor. In this perspective, chat-oriented dialogue systems are attracting the increasing attention of both research and practitioners interested in the simulation of natural dialogues with embodied conversational agents (Klüwer, 2011), conversational interfaces for smart devices (McTear et al., 2016) and the Internet of Things (Kar and Haldar, 2016). As a consequence, we are assisting to the flourishing of dedicated research venues on chat-oriented interaction. It is the case of WOCHAT[1], the Special Session on Chatbots and Conversational Agents, now at its second edition, as well as the Natural Language Generation for Dialogue Systems special session[2], both co-located with the Annual

[1] http://workshop.colips.org/wochat/@sigdial2017/

[2] https://sites.google.com/view/nlg4ds2017

SIGdial Meeting on Discourse and Dialogue.

While not representing any deep understanding of the interaction dynamics, speech acts can be successfully employed as a coding standard for natural dialogues tasks. In this report, we describe the first edition of the "itaLIan Speech acT labEliNg" (iLISTEN) task at the EVALITA 2018 campaign (Caselli et al., 2018). Among the various challenges posed by the problem of enabling conversational access to information, this shared task tackles the problem of recognition of the illocutionary force, i.e. the speech act, of a dialogue turn, that is the communicative goal of the speaker.

The remainder of the paper is organized as follows. We start by explaining the task in Section 2. In Section 3, we provide a detailed description of the dataset of dialogues, the annotation schema, and the data format and distribution protocol. Then, we report about the evaluation methodology (see Section 4) and describe the participating systems and their performance (see Section 5). We provide final remarks in Section 6.

2 Task Description

The task consists in automatically annotating dialogue turns with *speech act labels*, i.e. with the communicative intention of the speaker, such as statement, request for information, agreement, opinion expression, general answer, etc. Table 1 reports the full set of speech act labels used for the classification task, with definition, examples, and distribution in our corpus. Regarding the evaluation procedure, we assess the ability of each system to issue the correct speech act label among those included in the taxonomy used for annotation, described in the Section 3. Please, note that the participating systems are requested to issue labels only for the speech act used for labeling the user's dialogue turns, as futher detailed in the following.

3 Development and Test Data

3.1 A Dataset of Dialogues

We leverage the corpus of natural language dialogues collected in the scope of previous research about interaction with Embodied Conversational Agents (ECAs) (Clarizio et al., 2006), in order to speed up the process of building a gold standard. The corpus contains overall transcripts of 60 dialogues, 1,576 user dialogue turns, 1,611 system turns and about 22,000 words.

The dialogues were collected using a Wizard of Oz tool as dialogue manager. Sixty subjects (aged between 21–28) were involved in the study, in two interaction mode conditions: thirty of them interacted with the system in a written-input setting, using keyboard and mouse; the remaining thirty dialogues were collected with users interacting with the ECA in a spoken-input condition. The dialogues collected using the spoken interaction mode were manually transcribed based on audio-recording of the dialogue sessions.

During the interaction, the ECA played the role of an artificial therapist and the users were free to interact with it in natural language, without any particular constraint: they could simply answer the question of the agent or taking the initiative and ask questions in their turn, make comments about the agent behavior or competence, argument in favor or against the agent's suggestion or persuasion attempts. The Wizard, on his behalf, had to choose among a set of about 80 predefined possible system moves. As such, the system moves (see Table 2) are provided only as a context information but are not subject to evaluation and do not contribute to the final ranking of the participant systems. Conversely, the participating systems are evaluated on the basis of the performance observed for the user dialogue turns (see Table 1).

3.2 Annotation Schema

Speech acts can be identified with the communicative goal of a given utterance, i.e. it represents its meaning at the level of its *illocutionary force* (Austin, 1962). In defining dialogue act taxonomies, researchers have been trying to solve the trade-off between the need for formal semantics and the need for computational feasibility, also taking into account the specificity of the many application domains that have been investigated (see (Traum, 2000) for an exhaustive overview). The Dialogue Act Markup in Several Layers (DAMSL) represents an attempt by (Core and Allen, 1997) to define a domain independent framework for speech act annotation.

Defining a speech act markup language is out of the scope of the present study. Therefore, we adopt the original annotation of the Italian advice-giving dialogues. Table 1 shows the set of nine labels employed for the purpose of this study, with definitions and examples. These labels are used for the annotation of the users' dialogue turns and are the object of classification for this task. In ad-

Table 1: The set of *user* speech act labels employed in our annotation schema. The participating systems are required to issue a label for the user moves only.

Speech Act	**Description**	**Example**	**Freq.**
OPENING	Dialogue opening or self-introduction	*'Ciao, io sono Antonella'*	2%
CLOSING	Dialogue closing, e.g. farewell, wishes, intention to close the conversation	*'Va bene, ci vediamo prossimamente'*	2%
INFO-REQUEST	Utterances that are pragmatically, semantically, and syntactically questions	*'E cosa mi dici delle vitamine?'*	25%
SOLICIT-REQ-CLARIF	Request for clarification (please explain) or solicitation of system reaction	*'Mmm, si ma in che senso?'*	7%
STATEMENT	Descriptive, narrative, personal statements	*'Penso che dovrei controllare maggiormente il consumo di dolciumi.'*	33%
GENERIC-ANSWER	Generic answer	*'Si'*, *'No'*, *'Non so.'*	10%
AGREE-ACCEPT	Expression of agreement, e.g. acceptance of a proposal, plan or opinion	*'Si, so che è importante.'*	5%
REJECT	Expression of disagreement, e.g. rejection of a proposal, plan, or opinion	*'Ho sentito tesi contrastanti al proposito.'*	5%
KIND-ATT-SMALLTALK	Expression of kind attitude through politeness, e.g. thanking, apologizing or smalltalk	*'Thank you.'*, *'Sei per caso offesa per qualcosa che ho detto?'*	11%

dition, in Table 1 we report the speech act labels used for the dialogue moves of the system, i.e. the conversational agent playing the role of the artificial therapist. The speech act taxonomy refines the DAMSL categories to allow appropriate tagging of the communicative intention with respect to the application domain, i.e. persuasion dialogues in the healthy eating domain.

In Table 3 we provide an excerpt from a dialogue from our gold standard. The system moves (dialogue moves and corresponding speech act labels) are chosen from a set of predefined dialogue moves that can be played by the ECA. As such, they are not interesting for the evaluation and ranking of participating systems and are provided only as contextual information. Conversely, the final ranking of the participating systems is based on the performance observed only on the prediction of speech acts for the users' move, with respect to the set of labels provided in Table 1. Please, note that the two sets of speech act labels for the user and the system moves, in Table 1 and Table 2, respectively, only partially overlap. This is due to the fact that the set of agent's moves includes also speech acts (such as persuasion attempts) that are observed only for the agent, given its caregiver role in the dialogue systems. Vice versa, some speech act labels (such as clarification questions) are relevant only for the user moves.

3.3 Data Format and Distribution

We provide both the training and testing dialogues in the XML format following the structure proposed in Figure 1. Each participating initially had access to the training data only. Later, the unlabeled test data were released during the evaluation period. The development and test data set contain 40 and 20 dialogues, respectively, equally distributed with respect to the interaction mode (text- vs. speech-based interaction).

4 Evaluation

Regarding the evaluation procedure, we assess the ability of each system to issue the correct speech act label for the user moves. The speech act label used for annotation of the user moves are reported in Table 1.

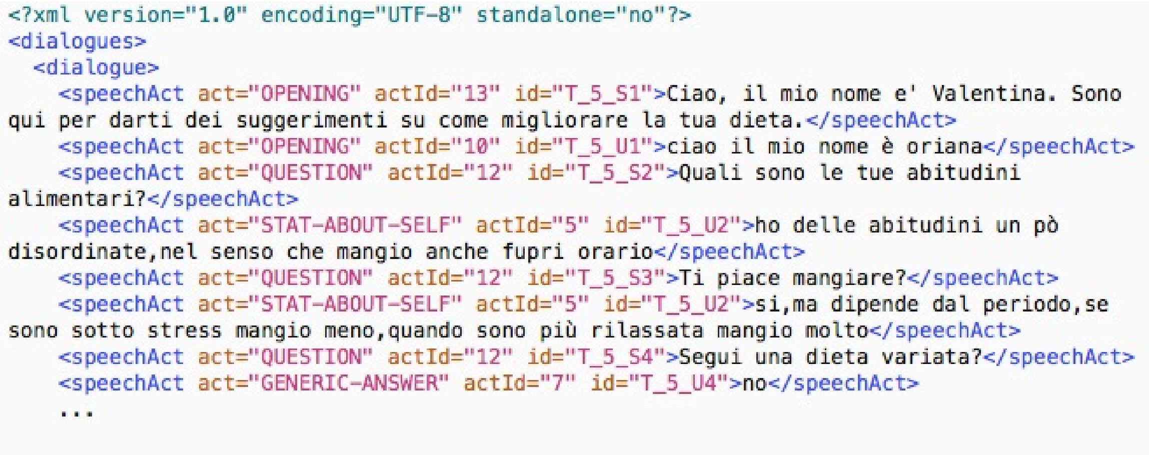
Specifically, we compute precision, recall and F1-score (macroaveraging) with respect to our gold standard. This approach, while more verbose than a simple accuracy test, arise from the need to correctly address the unbalanced distribution of labels in the dataset. Furthermore, by providing detailed performance metrics, we intend to enhance interesting discussion on the nature of the problem and the data, as they might emerge from the participants' final reports. As a baseline, we use the most frequent label for the user speech acts (i.e., STATEMENT).

Table 2: The set of *system* speech act labels in our annotation schema. These labels are provided as context information, i.e. the participating systems are *not* required to issue a label for the system moves.

Speech Act	Description	Example	Freq.
OPENING	Initial self-introduction by the ECA	*'Ciao, il mio nome è Valentina e sono qui per darti suggerimenti su come migliorare la tua dieta.'*	4%
CLOSING	Dialogue closing, e.g. farewell, wishes, intention to close the conversation	*'Grazie e arrivederci!'*	4%
QUESTION	Question about the user eating habits or information interests	'...'	32%
TALK-ABOUT-SELF	Statement describing own abilities, role and skills	*'Non sono in grado di mangiare, e quindi non seguo diete particolari.'*	3%
ANSWER	Provision of generic information after a user request	*'Purtroppo sono poco esperta in questo campo.'*	17%
OFFER-GIVE-INFO	Generic offer of help or specific information about the diet domain.	*'Idealmente un pasto dovrebbe contenere da 75 a 100 grammi di carne, un frutto di media grandezza, un piccolo piatto di pasta o 50 grammi di gelato. I cibi pronti permettono di controllare le porzioni riportando sulla confezione il valore energetico in calorie.'*	8%
PERSUASION-SUGGEST	Persuasion attempt performed by the conversational agent about dieting.	*'Lo so: mangiare, e forse vivere, in modo disordinato, sembra a molti piacevole,e forse hanno ragione. Ma a lungo andare gli effetti negativi si fanno sentire.'*	24%
ENCOURAGE-SORRY	Statements aimed at enhancing the user motivation or to apologize.	*'Questa e' una buona idea! '*	9%

Table 3: An excerpt from one of the dialogues in our dataset.

Actor	ID	Speech Act	Dialogue Move
SYSTEM	5-S1	OPENING	Ciao, il mio nome è Valentina. Sono qui per darti dei suggerimenti su come migliorare la tua dieta
USER	5-U1	OPENING	Ciao, il mio nome è Oriana.
SYSTEM	5-S2	INFO-REQUEST	Quali sono le tue abitudini alimentari?
USER	5-U2	STATEMENT	Ho delle abitudini disordinate, mangio anche fuori orario.
SYSTEM	5-S3	INFO-REQUEST	Ti piace mangiare?
USER	5-U3	STATEMENT	Si, ma dipende dal periodo: se sono sotto stress mangio meno, quando sono più rilassata mangio molto
SYSTEM	5-S4	INFO-REQUEST	Segui una dieta variata?
USER	5-U3	GENERIC-ANSWER	No.



```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<dialogues>
  <dialogue>
    <speechAct act="OPENING" actId="13" id="T_5_S1">Ciao, il mio nome e' Valentina. Sono
qui per darti dei suggerimenti su come migliorare la tua dieta.</speechAct>
    <speechAct act="OPENING" actId="10" id="T_5_U1">ciao il mio nome è oriana</speechAct>
    <speechAct act="QUESTION" actId="12" id="T_5_S2">Quali sono le tue abitudini
alimentari?</speechAct>
    <speechAct act="STAT-ABOUT-SELF" actId="5" id="T_5_U2">ho delle abitudini un pò
disordinate,nel senso che mangio anche fupri orario</speechAct>
    <speechAct act="QUESTION" actId="12" id="T_5_S3">Ti piace mangiare?</speechAct>
    <speechAct act="STAT-ABOUT-SELF" actId="5" id="T_5_U2">si,ma dipende dal periodo,se
sono sotto stress mangio meno,quando sono più rilassata mangio molto</speechAct>
    <speechAct act="QUESTION" actId="12" id="T_5_S4">Segui una dieta variata?</speechAct>
    <speechAct act="GENERIC-ANSWER" actId="7" id="T_5_U4">no</speechAct>
    ...
```

Figure 1: Data format

Table 4: Overall micro- and macro-averaged Precision, Recall, and F-score for the participating systems

	Micro			***Macro***		
System	**Prec**	**Rec**	**F**	**Prec**	**Rec**	**F**
UNITOR.kelp	0.7328	0.7328	0.7328	0.6810	0.6274	0.6531
X2Check.c2c	0.6848	0.6848	0.6848	0.6076	0.5844	0.5957
Baseline	0.3403	0.3403	0.3403	0.0378	0.1111	0.0564

Table 5: Precision, Recall, and F-score values by speech act labels

	Unitor			*X2Check*		
Class	**Prec**	**Rec**	**F**	**Prec**	**Rec**	**F**
OPENING	1.0000	1.0000	1.0000	1.0000	0.7273	0.8421
CLOSING	0.7778	0.7000	0.7368	0.8182	0.9000	0.8571
INFO-REQUEST	0.7750	0.8304	0.8017	0.7355	0.7946	0.7639
SOLICITATION-REQ-CLARIF	0.4000	0.3333	0.3636	0.4444	0.3333	0.3810
STATEMENT	0.7500	0.9444	0.8361	0.6667	0.8957	0.7644
GENERIC-ANSWER	0.8571	0.9231	0.8889	0.7581	0.9038	0.8246
AGREE-ACCEPT	0.6471	0.4583	0.5366	0.5714	0.5000	0.5333
REJECT	0.4286	0.0769	0.1304	0.0000	0.0000	0.0000
KIND-ATT-SMALLTALK	0.5000	0.3864	0.4359	0.4737	0.2045	0.2857

5 Participants and Results

The task was open to everyone from industry and academia. Sixteen participants registered, but only two teams actually submitted the results for the evaluation. A short description of each system follows:

UNITOR - The system described in (Croce and Basili, 2018) is a supervised system which relies on a Structured Kernel-based Support Vector Machine for making the classification of the dialogue turns sensitive to the syntactic and semantic information of each utterance. The Structured Kernel is a Smoothed Partial Tree Kernel (Croce et al., 2011) that exploits both the parse tree and the cosine similarity between the word vectors in a distributional semantics model. The authors use the tree parser provided by SpaCy[3] and the Kelp framework[4] for SVM.

X2Check - The team did not submit the report.

The performance of the participating systems is evaluated based on the macro (and micro) precision and recall (Sebastiani, 2002). However, the official task measure used to rank the systems is the macro-F. Results are reported in Table 4.

The best performance (0.6531) is provided by the UNITOR system. Both systems are able to overcome the baseline also for micro-F. The baseline has a low macro-F since it predicts always the same class (STATEMENT) and for the other classes the F-measure is zero. As expected, the micro-F overcomes the macro-F since some classes are hard to predict due to the low number of examples in the training data, such as AGREE, SOLICITATION-REQ-CLARIF and REJECT. Precision, Recall, and F-score values by speech act labels are showed in Table 5.

We also provide the confusion matrix for each system, respectively Table 6 for UNITOR and Table 7 for X2Check. We observe that, for both systems, the class REJECT is the most difficult to classify. This evidence is consistent with the findings from previous research on the same corpus of dialogues (Novielli and Strapparava, 2011). In particular, we observe that dialogue moves belonging to the REJECT class are often misclassified as STATEMENT. More in general, the main cause of error is the misclassification as STATEMENT. One possible reason is that statements represent the majority class, thus inducing a bias in the classifiers. Another possible explanation, is that dialogue moves that appear to be linguistically consistent with the typical structure of statements have been annotated differently, according to the actual communicative role they play.

[3] `https://spacy.io/`

[4] KeLP is a Java Kernel-based Learning Platform: `http://www.kelp-ml.org/`

Table 6: Confusion Matrix of the UNITOR system w.r.t. gold standard. In column the number of classes from the gold standard, while rows report the system decisions. In bold correct classifications.

	STATEMENT	KIND-ATT.	GEN.-ANSW.	REJECT	CLOSING	SOL.-CLAR.	OPENING	AGREE	INFO-REQ.
STATEMENT	**153**	6	3	24	0	3	0	2	13
KIND-ATT.	4	**17**	0	5	1	2	0	3	2
GEN.-ANSW.	1	0	**48**	0	0	1	0	6	0
REJECT	0	3	0	**3**	0	0	0	0	1
CLOSING	0	0	0	0	**7**	1	0	1	0
SOL.-CLAR.	0	6	0	2	1	**8**	0	1	2
OPENING	0	0	0	0	0	0	**11**	0	0
AGREE	0	3	1	1	0	0	0	**11**	1
INFO-REQ.	4	9	0	4	1	9	0	0	**93**

Table 7: Confusion Matrix of the X2Check system w.r.t. gold standard. In column the number of classes from the gold standard, while rows report the system decisions. In bold correct classifications.

	STATEMENT	KIND-ATT.	GEN.-ANSW.	REJECT	CLOSING	SOL.-CLAR.	OPENING	AGREE	INFO-REQ.
STATEMENT	**146**	15	3	30	1	2	1	2	19
KIND-ATT.	2	**9**	0	0	0	1	0	5	2
GEN.-ANSW.	5	3	**47**	2	0	3	0	2	0
REJECT	0	0	0	0	0	0	0	0	0
CLOSING	0	0	0	1	**9**	0	0	1	0
SOL.-CLAR.	1	4	0	2	0	**8**	1	0	2
OPENING	0	0	0	0	0	0	**8**	0	0
AGREE	2	5	1	0	0	1	0	**12**	0
INFO-REQ.	7	8	1	4	0	9	1	2	**89**

6 Final Remarks and Conclusions

We presented the first edition of the new shared task about itaLIan Speech acT labEliNg (iLISTEN) at EVALITA 2018. The task fits in the fast-growing research trend focusing on conversational access to the information, e.g. using chatbots or conversational agents. The task consists in automatically annotating dialogue turns with speech act labels, representing the communicative intention of the speaker. The corpus of dialogues has been collected in the scope of previous research on natural language interaction with embodied conversational agents. Specifically, the participating systems had to annotate the speech acts associated to the user dialogue moves while the agent's dialogue turns were provided as context.

We received two runs from two teams, one from academia and the other one from industry. In spite of the inherent complexity of the tasks, both systems largely outperformed the baseline, represented by the trivial classifier always predicting the majority class for users' moves. The best performing system leverages syntactic features and relies on a Structured Kernel-based Support Vector Machine. Follow-up editions might involve extending the benchmark with dialogues from different domains. Similarly, dialogues in different languages might be also included in the gold standard, as done for Automatic Misogyny Identification task at EVALITA 2018 (Fersini et al., 2018). This would enable to assess to what extent the task is inherently dependent on the language and how the proposed approaches are able to generalize.

References

John L. Austin. 1962. *How to do things with words.* William James Lectures. Oxford University Press.

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. EVALITA 2018: Overview of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Giuseppe Clarizio, Irene Mazzotta, Nicole Novielli, and Fiorella De Rosis. 2006. Social attitude towards a conversational character. pages 2–7.

Mark G. Core and James F. Allen. 1997. Coding Dialogs with the DAMSL Annotation Scheme.

Danilo Croce and Roberto Basili. 2018. A Markovian Kernel-based Approach for itaLIan Speech acT labEliNg. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2011. Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of EMNLP*.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and

Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Rohan Kar and Rishin Haldar. 2016. Applying Chatbots to the Internet of Things: Opportunities and Architectural Elements. *CoRR*, abs/1611.03799.

Tina Klüwer. 2011. "I Like Your Shirt" - Dialogue Acts for Enabling Social Talk in Conversational Agents. In *Intelligent Virtual Agents*, pages 14–27.

Michael McTear, Zoraida Callejas, and David Griol Barres. 2016. *The Conversational Interface: Talking to Smart Devices*. Springer International Publishing.

Nicole Novielli and Carlo Strapparava. 2011. Dialogue act classification exploiting lexical semantics. In *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*, chapter 4, pages 80–106. IGI Global.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, London.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Comput. Linguist.*, 26(3):339–373, September.

David R. Traum. 2000. 20 Questions for Dialogue Act Taxonomies. *Journal of Semantics*, 17(1):7–30.

Soroush Vosoughi and Deb Roy. 2016. A Semi-automatic Method for Efficient Detection of Stories on Social Media. In *Proc. of the 10th AAAI Conf. on Weblogs and Social Media.*, ICWSM 2016, pages 711–714.