A topic recognition system for real world human-robot conversations

S.Anzalone, Y.Yoshikawa, H.Ishiguro, E. Menegatti, E. Pagello, R. Sorbello

Abstract

One of the main features of social robots is the ability to communicate and interact with people as partners in a natural way. However, achieving a good verbal interaction is a hard task due to the errors on speech recognition systems, and due to the understanting the natural language itself. This paper tries to overcome such kind of problems by presenting a system that enables social robots to get involved in conversation by recognizing its topic. Through the use of classical text mining approach, the presented system allows social robots to understand topics of conversation between human partners, enabling the customization of behaviours in their accordance. The system has been evaluated in different contexts, taking in account the quality and accuracy of the speech recognition system used by the social robot.

S.Anzalone · Y.Yoshikawa · H.Ishiguro
Intelligent Robotics Laboratory
Department of Systems Innovation
Graduate School of Engineering Science
Osaka University

E. Menegatti · E. Pagello
Intelligent Autonomous Systems Laboratory (IAS-Lab)
Department of Information Engineering
Faculty of Engineering
University of Padua

R. Sorbello
Robotics Lab
Department of Computer Engineering
Faculty of Engineering
University of Palermo

Fig. 1 A sketch of the working environment of the system.

1 Introduction

Social robots are system able to communicate and interact as real partners with humans [1]. Communication between humans can be classified in two different kinds: verbal and non-verbal. While non-verbal communication is based on gazing, pointing, gesturing or changing of facial expressions, verbal communication is fully based on the speech [2]. Consequently, human speech is a natural and intuitive interface for communicating with robot. Despite of this, it is very difficult to achive good interaction with social robots using verbal communication because the nature of the auditive patterns and the nature of the human speech itself [3]. The auditive flow contains a lot of information that is hard to manage: environmental conditions with noise and echos, are the first problem to deal; a more complex matter to achieve is focusing the attention of the system on a single talker among a mix of several conversation and background noise, problem yet described in literature as “cocktail party effect”. At last, the human speech itself encodes several kind of information: who is the speaker, the speaker’s identity; what the speaker is saying, the speaker’s speech; how the speaker said it, the speaker’s prosody.

Beyond these problems, that are all relative to the auditory recogition of the speech, another issue concerns the human natural language itself [4]: linguists do not have a complete understanding of the underlying rules of spoken languages because it seems impossible to describe them only in terms of syntax, semantics or phonetics rules, as it is possible to constructed language. The comprehension of the real meaning of the speech becomes a very hard task due to its incompleteness, ambiguity and semi structured or unstructured characterization.

Several studies tried to deal with these obstacles from both the speech recongition side and from the natural language processing point of view. Researchs tried to improve the accuracy of the speech recognition systems using a more detailed description of phonemes or through triphones [5] or using larger vocabulary or by providing additional constraints [6]. On the other side, researchs on natural language processing tried to achieve a deep understanding of the spoken utterances improving parsers, using stochastic models [7], context based ontologies or rich lexical databases that includes semantic information, such as Wordnet [8].

From the point of view of the researches on verbal communication for social robots many attempts have been performed. Simple command based systems gave

important results, but users should knows the commands or should be previously instructed about the behaviour of the robot [9]. Dialogue based systems have also been successfully used [10], but also in this case it is very difficult to achieve a free conversation speech, due to the dialogue system itself, that should be able to cover a wide spread of possible conversation paths. Some systems tried to use ontologies to retrieve a complete understanding of the conversation [11]. Other social robot systems tried to use customized algorithms, such as Latent Semantic Analysis, to extract from the speech some important characterizations, as well as the emotions [12] [13]. All these systems soffer from the problem of having errors on the input speech utterance, due to mistake of the speech recognition system[14].

The system presented in this paper allows social robots to get involved in conversation with humans as shown in Figure 1, by recognizing the topic of the current conversation. The system will try to overcome the low recognition rate on the accuracy of the speech recognition system by grounding conversation between people to its topic, using only the relevant words.

2 System overview

The assumption made by this work is that it is difficult to obtain accurate and correct results from a speech recognition systems in real world applications. In order to avoid this problem, the system presented in this paper will recognize topics of the conversation in which a robot si involved by ignoring the details of the structure of each sentence, focusing only to the important words that will reveal what the the people is talking about.

The scenario considered by this system is a low-noise environment in which, as shown in Figure 1, two (or more) human partners talk in turn about a closed set of topics. The system has been supposed to be an efficent tool for several kind of application, such as to interact in a customized way as robotic companion, or to suggest information related to what people is talking about, as a robotic assistant, or as system able to profile human partners according to their favourite topics.

The system is composed by several reusable modules capable to cooperate and exchange information, as shown in Figure 2. The framework ROS allowed the development of these components to distribute the algorithms through several software units following a modular top-down approach. In particular, two main modules have been implemented: the speech recognition module and the topic classification module.

The speech recognition module is implemented through the use of Julius[15]. Julius is a state-of-art large vocabulary continuous speech recognition system that is able to perform in realtime. The system has been developed in a context of Japanese language users, then Julius has been traineed using a Japanese language model composed by 20k words from newspaper articles and an acustic model based on triphones to assure efficency and high performances.

The topic classification module carries out the recognition of the topic that emerges from the speech utterance decoded by using a slightly modified version of the "Term Frequency - Inverse Document Frequency" (TF-IDF) weighting, called "Term Frequency - Inverse Topic Frequency" (TF-ITF). The output of this system will be the set of probabilities related to each possible topic.

3 Topic classification

The topic recognition system is based on a modified version of the TF-IDF ranking function, often used in text mining and information retrival. Given a corpus of documents, the TF-IDF weight is a statistical measure that evaluates how much a word is important to a document. The apporach chosen was to evaluate the TF-IDF weight for each word in a corpus of documents related to several topics by calculating how much a word is relevant to each topic, rather than each document, realizing a "Term Frequency - Inverse Topic Frequency" instead of the classical "Term Frequency - Inverse Document Frequency".

Given a corpus of documents labeled by their topic, the weight of the word W according to a topic T is calculated by the formula:

$$TF\text{-}ITF(W,T) = \frac{freq(W,T)}{\sum_w freq(w,T)} \times \log \frac{\sum_w \sum_{\substack{t \\ t \neq T}} freq(w,t)}{\sum_{\substack{t \\ t \neq T}} freq(W,t)}$$

The first part of the formula describe the term W using its frequency normalized by the number of the words of the topic T. The important property of the normalized frequency is that some words are more used in some contexts rather in others. Despite of this, the term frequency weight is not enough because many words, such as particles or auxiliar verbs, are used a lot in every context. The second part of the formula tries to overcome this limit, penalizing the terms that are used in the other topics.

According to this approach, it is possible to recognize and discard all the negligible terms by applying a simple thresholding to the TF-ITF: words with a higher TF-ITF weight in a cosidered topic will be taken in account as meaningful for that topic.

After a normalization of the TF-ITF weight of each word by its weight in all the topics, it is possible to analyze and classify complex sentences through the formula:

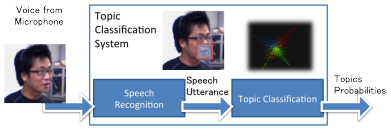

Fig. 2 A sketch of the ROS modules that compose the system.

$$P(S,T) = 1 - \prod_{\forall w \in S} \left[1 - TF\text{-}ITF(w,T)\right]$$

Through a sequence of products the TF-ITF weight of each word inside a sentence is used to obtain the sentence probability to belong to a class T. For a given sentence, the topic classification system will in this way produce the probabilities for each trained topic.

4 Experimental results

The system has been trained to recognize four different topics: “soccer”, “ski”, “baseball” and “swimming”. For each topic, a folder of documents has been selected from the Japanese Wikipedia pages. Documents have been chosen in order to have about 20000 words for each category, equally distributed among the topics. The TF-IDF weight has been calculated for each word according to each of the four category, then, experimentally, a threshold for distinguish meaningless words, such as auxiliar verbs, particles, and stopwords, has been found.

Tests of the system have been performed in three kind of situation, as shown in Figure 3: using raw text from sport newspapers, without the use of the speech recognition system; using read speech in a controlled environment, through japanese television sport newscasts captured from YouTube; using spontaneous speech, captured in real conversation between people.

Experiments achieved have been evaluated in terms of accuracy, precision, sensitivity and specificity.

4.1 Raw text

Raw text has been collected from japanese sport news websites. For each category a set of 10 documents have been chosen to evaluate their main topic. In this case the results obtained are free from the speech recogntion errors because the raw text

Fig. 3 Raw text, read text and free speech conversation.

is submitted directly to the topic classification system. Results shown the performances of the classification system in an ideal scenario using an errorless speech recognition system.

		Predicted class			
		Soccer	Baseball	Ski	Swim
Actual class	Soccer	100	0	0	0
	Baseball	0	100	0	0
	Ski	0	0	100	0
	Swim	0	10	10	80

Table 1 The confusion matrices related to the newspapers raw text classification. Data is expressed in percentage.

Confusion matrix (see Table 1) and performances of the system shown an high reliability of topic classification system. In particular, experiments reported 98% of accuracy, 95% of precision, 97% of sensitivity, 98% of specificity, as shown in Table 4.

4.2 Read text

For each category, five videos from YouTube have been collected. Videos of about one minute of length have been chosen from japanese television sport newscasts in order to evaluate the performances of the whole system, that now includes also the speech recognition system, in a controlled, noiseless, environment. Moreover, the use of newcasts videos allows the evaluation of the system in a best-case scenario because anchormen will talk using a formal diction.

		Predicted class			
		Soccer	Baseball	Ski	Swim
Actual class	Soccer	60	0	20	20
	Baseball	0	100	0	0
	Ski	0	0	100	0
	Swim	0	0	0	100

Table 2 The confusion matrices related to the read text classification. Data is expressed in percentage.

Due to the use of the speech recognition system, a recognition rate performance has been added to the measures collected. As shown in the confusion matrix in Table 2, performances are still high, despite the recognition rate of the speech recognition system of 25%. Experiments shown 93% of accuracy, 90% of precision, 88% of sensitivity, 96% of specificity, as depicted in Table 4.

4.3 Spontaneous speech

To obtain spontaneous speech samples, six persons have been involved in an experiment. A video for each of the four categories has been shown to each person, then 3 couples have been formed. Each partecipant has been invited to describe and conversate in turn about a single video, for about one minute and half. Experiments have collected six audio samples for each of the four categories, a total of 24 audio samples of conversation. Spontaneous speech data set collected has been evaluated by the use of the whole system.

		Predicted class			
		Soccer	Baseball	Ski	Swim
Actual class	Soccer	100	0	0	0
	Baseball	0	66	33	0
	Ski	16	0	66	16
	Swim	16	0	0	83

Table 3 The confusion matrices related to the spontaneous speech classification. Data is expressed in percentage.

As it is possible to see in Table 3, the low recognition rate of the speech recognition system of 13% affects the performances of the whole system. However, despite of this, results are still very significative because their reliability. Experiments obtained 89% of accuracy, 79% of precision, 81% of sensitivity, 93% of specificity, as in Table 4.

4.4 Results comparison and limitations

The Table 4 shown a comparison between the different scenarios in which the system has been tested. Despite of the recognition rate of the speech recognition system, performances of the classification system are still high. This is a direct effect obtained by taking in account only some words classified as important in the considered contexts, while forgetting about the details: the speech recognition system does not need to be extremly accurate.

While achieving these performances, the system incur into several limitation. Independence of the topics to be recognized can affect the performances of the system, due to the underlied naive independence assumption. While the system is able to distinguish highly uncorrelated, independent classes, such as “baseball” and “swimming”, more errors may occour while trying to separate more related topics, such as “swimming” and “water polo”. In this case, many words can be used in both the topic considered, such as “water” or “pool”, reducing the reliability of the final results. At last, as shown in Figure 4, tests performed with chunks of conversation

	Raw Text	Read Text	Spontaneous Speech
Recognized Speech	-	25	13
Accuracy	98	93	89
Precision	95	90	79
Sensitivity	97	88	81
Specificity	98	96	93

Table 4 Performances comparison in different scenarios. Data is expressed in percentage.

with different lenght shown that the system is able to assess the topic in the 80% of the experiments during their first 20 seconds.

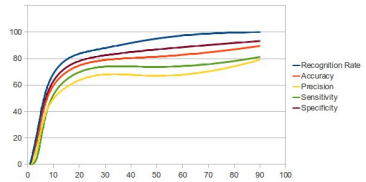

Fig. 4 Spontaneous speech performances among the time [sec].

5 Conclusion and future work

A natural language processing system for social robots involved in human conversations has been developed and tested. The system was based upon a slightly modified version of "Term Frequency - Inverse Document Frequency" weighting called "Term Frequency - Inverse Term Frequency", that allowed the system to forget about the details of the sentences, while recognizing the topic of the current conversation occurring between human partners. Experiments of the system performed in several scenarios shown the benefits and the limits of the presented approach.

Results shown encourage to pursuit on the experimentation of this approach in new, real, more complicated scenarios. While the presented system used Julius as state of the art speech recognition system, more experiments should be performed by using different and more efficent systems. Moreover, in order to assure better performances, a sound localization system will be used to try to deal with noisy environment and cocktail party effects. New experiments will try to capture relevant information from vocal interaction between robot's human partners in order to adapt its behaviours according to the occurring conversation, to profile them or to simply suggest them conversation related information.

Acknowledgments

Authors would like to thank Dr. S. Livieri, Dr. F. Dalla Libera, Prof. A. Chella, Prof. G. Vassallo for their support during the development of this project.

This work has been supported by a Grant-in-Aid for scientific research fellowships from Japan Society for the Promotion of Science (JSPS). This project was partially founded by Regione Veneto Cod. progetto: 2105/201/5/2215/2009 approved with DGR n. 2215, 21/07/2009.

References

1. C. Breazeal. Toward sociable robots. *Robotics and Autonomous Systems*, 42(3-4), 2003.
2. C.L. Breazeal. *Designing sociable robots*. The MIT Press, 2004.
3. J. Benesty, M.M. Sondhi, Y. Huang, and S. Greenberg. Springer handbook of speech processing. *The Journal of the Acoustical Society of America*, 126:2130, 2009.
4. P. Jackson and I. Moulinier. *Natural language processing for online applications: Text retrieval, extraction and categorization*, volume 5. John Benjamins Pub Co, 2007.
5. S. Darjaa, M. Cerňak, M. Trnka, M. Rusko, and R. Sabo. Effective triphone mapping for acoustic modeling in speech recognition. In *Proceedings of Interspeech 2011 Conference*, 2011.
6. A. Lee and T. Kawahara. Recent development of open-source speech recognition engine julius. In *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, pages 131–137. Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee, 2009.
7. Y.W. Teh. Bayesian tools for natural language learning. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 219–219. Association for Computational Linguistics, 2011.
8. C. Fellbaum. Wordnet. *Theory and Applications of Ontology: Computer Applications*, 2010.
9. B.K. Shim, Y.K. Cho, J.B. Won, and S.H. Han. A study on real-time control of mobile robot with based on voice command. In *Control, Automation and Systems (ICCAS), 2011 11th International Conference on*, pages 1102–1103. IEEE, 2011.
10. I. Toptsis, S. Li, B. Wrede, and G.A. Fink. A multi-modal dialog system for a mobile robot. In *Eighth International Conference on Spoken Language Processing*, 2004.
11. S. Kobayashi, S. Tamagawa, T. Morita, and T. Yamaguchi. Intelligent humanoid robot with japanese wikipedia ontology and robot action ontology. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 417–424. ACM, 2011.
12. SM Anzalone, F. Cinquegrani, R. Sorbello, and A. Chella. An emotional humanoid partner. *Linguistic and Cognitive Approaches To Dialog Agents (LaCATODA 2010) At AISB*, 2010.
13. A. Chella, R. Sorbello, G. Pilato, G. Vassallo, G. Balistreri, and M. Giardina. An architecture with a mobile phone interface for the interaction of a human with a humanoid robot expressing emotions and personality. *AI*IA 2011: Artificial Intelligence Around Man and Beyond*, pages 117–126, 2011.
14. F. Kraft, K. Kilgour, R. Saam, S. Stuker, M. Wolfel, T. Asfour, and A. Waibel. Towards social integration of humanoid robots by conversational concept learning. In *Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on*, pages 352–357. IEEE, 2010.
15. A. Lee, T. Kawahara, and K. Shikano. Julius—an open source real-time large vocabulary recognition engine. In *Seventh European Conference on Speech Communication and Technology*, 2001.