Automatic Identification of Speakers' Intentions in A Multi-Party Dialogue System

Vasile Rus, Cristian Moldovan, Amy Witherspoon, Arthur C. Graesser
The University of Memphis

We focus in this paper on automatic identification of speakers' intentions in multi-party dialogue systems. Speakers' intentions are modeled using elements from the speech act theory (Austin, 1962; Searle, 1969) which was developed based on the "language as action" assumption. The language as action assumption implies that when people say something, they do something (Austin, 1962).

According to Searle (1969), there are three levels of action carried by language in parallel: first, there is the locutionary act which consists of the actual utterance and its exterior meaning; then, there is the illocutionary act, which is the real intended meaning of the utterance, its semantic force; finally, there is the perlocutionary act which is the actual effect of the utterance, such as scaring, persuading, encouraging, etc.

The notion of speech act is closely linked to the illocutionary level of language. Usual illocutionary acts are: greeting (*"Hello, John!"*), asking questions (*"Is it snowing?"*), or making requests (*"Could you pass the salt?"*).

We hypothesized that the leading tokens in an utterance are indicative of the speaker's intention, i.e. speech act. For instance, a question most likely starts with a *wh*-word, such as *How*, followed by an auxiliary verb. In contrast, a statement starts with noun or pronoun followed by a verb.

Method

Dialogue Sample

The original data used for this analysis came from a study run using an epistemic game, *Urban Science*. Within the game, players take on the role of an intern for an Urban Planning company and are provided guidance from a mentor on the proper steps to be taken in redesigning a city. The players interacted with the mentor through a chat facility provided in the game. All chat among players and mentors was logged and the resulting data set included 1,956 mentor contributions and 2,175 player contributions.

A random sample of 750 mentor contributions and 750 player contributions was selected for hand annotation of the speech act categories. The 750 mentor and 750 student contributions were further split into speech acts, using end of sentence punctuation marks (i.e., periods, question marks, and exclamation marks) as delimiters. This process resulted in 901 mentor speech acts and 765 player speech acts. Each of these speech acts were hand annotated by one trained annotator. Prior to this annotation, two assistants trained on the speech act categories independently annotated 1,500 speech acts. The average inter-rater reliability (across all categories) was Kappa = 0.87. Each of the 1,666 mentor and player speech acts was annotated using only one of the categories from the speech act taxonomy. The speech act categories used are shown in the first column in Table 1. Examples and the distribution of the speech acts in the collected data set are also provided in Table 1.

Table 1. The set of speech act categories with examples and distribution.

Speech Act Category	Example	Count
Statement	I'll be your planning consultant.	396
Request	Click that and click "New Staff Page"	228
Reaction	Ah, I see.	287
MetaStatement	i didn't understand what maya wanted	104
Greeting	Hello!	66
ExpressiveEvaluation	good!!!!!!!!!!	103
Question	why am i getting notes from people not in my group?	415
Other	same thing what	67

Analytic Strategy

We adopted a supervised machine learning methodology in which we take our basic idea and map it onto a set of features. The features in our case are the leading tokens in an utterance. The feature space representation can be used with any machine learning algorithms to tune the parameters of the model according to expert-annotated data, i.e. training data. We have used two machine learning algorithms, Naïve Bayes and Decision Trees, to learn the parameters of the basic model and induce classifiers that can categorize new utterances into speech act categories. Naïve Bayes are statistical classifiers that make the naïve assumption of feature independence. While this assumption means models that are too simplistic at times, it helps with better estimating the parameters of the model which in turn leads to good classifiers in general. Decision Trees are based on the idea of organizing the features in a hierarchical decision tree based on information gain. More informative features are always higher in the tree. The accuracy of the induced classifiers was measured on separate testing data sets using 10-fold cross validation. Accuracy measures how well the predicted speech act categories match the correct categories, which were annotated by human experts. We experimented with n=2..8 leading tokens to make predictions about the speech act categories of the utterances (the average contribution has 7.26 tokens). If our hypothesis is correct then using 2 or 3 leading tokens should be at least as good as using 8 tokens or maybe better.

Results

A summary of our results is presented in Table 2. The columns indicate the number of leading/feature tokens used and a variety of performance measures such as accuracy, kappa statistics (a measure of agreement between predicted and expert categories that takes into consideration chance agreement), precision (average precision for all speech act categories; for a category, precision indicates the percentage of correct predictions out of all instances predicted as belonging to that category), recall (average recall for all classes; for one class, it indicates the percentage of true instances belonging to the category that were correctly predicted), and F-measure (harmonic mean of precision and recall). From the table, we can see that when using Naïve Bayes the best results obtained are for the model using n=3 leading tokens. For Decision

Trees, the best results are obtained for n=2. For larger n=4..8, the results are either worse (Naïve Bayes) or similar (Decision Trees). The same pattern is noticed for other performance measures as well. This supports our hypothesis that few leading words in an utterance are very predictive of its speech act category. In other words, a speaker's intentions can be identified with very good accuracy after hearing only a few of the words in her utterance.

Discussion

We provided in this paper evidence that the leading tokens of contributions in multi-part dialogue are indicative of the speakers' intentions coded as speech act categories. A post-hoc analysis of the results revealed that the classifiers were very good at correctly identifying all the speech act categories with no single class performing poorly. A closer look at the results revealed that when in error, Questions were most likely confused with Requests and Statements, which happens for indirect questions which look more or less like a Statement or for Requests that are stated in the form of a Question. Similarly, Greetings were mostly confused with Reactions which is understandable given that both are short. This is particularly true when using larger models that use 5 or 6 leading tokens to predict the speech act categories. For such large models, the artificial values we used to fill in the missing features make both Greetings and Requests look similar. For instance, when using the first 6 leading tokens to predict the speech act and the Greeting has one word as in *Hello!* (similarly, Reactions are very short as in *I see.*) then the remaining features are filled in with an artificial value such as *none*.

Our basic model has its own limitations, which explains the very good but not perfect performance results. We plan to extend the basic model we proposed here with more contextual clues, which we believe will lead to further improvements in performance. Contextual clues will exploit discourse sequential patterns that humans most likely take advantage of, such as the fact that after a greeting another greeting follows as a response to the first one.

Table 2. The distribution of speech act categories in our data set.

Number of feature tokens	Classification Algorithm	Accuracy	Kappa	Precision	Recall	F-Measure
N = 1	Naïve Bayes	66.0264 %	0.5727	0.666	0.66	0.638
	J48	66.2665 %	0.5832	0.671	0.663	0.653
N = 2	Naive Bayes	67.6471 %	0.5961	0.693	0.676	0.651
	J48	**68.3073 %**	0.6084	0.687	0.683	0.673
N = 3	**Naive Bayes**	**68.3073 %**	0.6051	0.693	0.683	0.655
	J48	67.9472 %	0.6041	0.686	0.679	0.67
N = 4	Naive Bayes	65.7863 %	0.5739	0.657	0.658	0.629
	J48	67.9472 %	0.604	0.686	0.679	0.67
N = 5	Naive Bayes	63.8055 %	0.5487	0.64	0.638	0.605
	J48	67.9472 %	0.604	0.686	0.679	0.67
N = 6	Naive Bayes	62.9652 %	0.5387	0.658	0.63	0.597
	J48	68.1273 %	0.6061	0.687	0.681	0.672
N = 7	Naïve Bayes	61.5246 %	0.5212	0.636	0.615	0.58
	J48	68.1873 %	0.6069	0.688	0.682	0.672
N = 8	Naïve Bayes	61.4046 %	0.5206	0.635	0.614	0.58
	J48	68.1273 %	0.6061	0.687	0.681	0.672

References

Austin, J.L. 1962. *How to do Things with Words.* Oxford University Press, 1962.

Searle, J.R. 1969. *Speech Acts.* Cambridge University Press, GB, 1969.

Witten, I. H.; and Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques, 2nd Edition.* Morgan Kaufmann, San Francisco.