УДК 004.912, 004.852

Классификатор диалоговых актов в интеллектуально-диалоговой системе

Быков И.А., ФГАОУ ВО СПбПУ, Санкт-Петербург
научный руководитель: Тимофеев Д.А., ФГАОУ ВО СПбПУ, Санкт-Петербург

В процессе реализации любой интеллектуально-диалоговой системы возникает задача классификации диалоговых реплик. Типичным решением данной проблемы является использование концепции интентов, где каждая входная реплика причисляется к одному фиксированному классу. В зависимости от полученного класса запускается обработчик, который работает по заранее определенному сценарию. Подобное решение, например, использует система iPavlov, которая позволяет вести целенаправленные диалоги с чат-ботами на заданные темы. Недостаток такого подхода заключается в том, что такие модели используют общий набор классов, которые применяются и для идентификации предметной области, и для распознавания сущностей, и для управления диалогами, что приводит к увеличению классов и снижению точности распознавания.

Существует другой более эффективный подход к построению классификатора, который использует классификацию пользовательских реплик на основе иерархии диалоговых актов. Различные диалоговые акты играют разные роли в управлении диалогом, что снимает ограничение, которое было существенным для интентов. Тем не менее доступные корпусы зачастую имеют несовместимую разметку, а значит исходные модели сильно привязаны к той области, в рамках которой они разрабатывались. Решением такой проблемы является использование разметки в соответствии с международным стандартом ISO 24617-2-2012, что позволит обеспечить единое формальное представление о каждой реплике пользователя без привязки к конкретной проблемно-ориентированной области. Существуют готовые корпусы размеченных диалогов в соответствии с данным стандартом, например DialogBank (DB), Common Alexa Prize Conversations (CAPC), Socialbot Logs (S-Logs). Они не привязаны к предметной области и соответственно могут быть использованы в разных системах. Однако такие корпусы достаточно малы и представлены в основном для английского языка.

В рамках описанной проблемы вытекает необходимость разработки алгоритма классификации диалоговых актов на русском языке. В качестве данных для обучения используются заранее подготовленные диалоги пользователей на определенные темы, а также тексты книг, пьес, опер. Ручная разметка такого корпуса в соответствии с описанным стандартом является достаточно трудоемкой задачей, поэтому итоговый корпус содержит небольшое количество данных. Существующие алгоритмы классификации диалоговых актов используют признаки, тесно связаны с грамматикой того языка, для которого они изначально разрабатывались. Для русского языка необходимо использовать другие признаки, так как с точки зрения грамматики язык существенно отличается от языков, для которых такие классификаторы доступны.

Для построения классификатора авторами предлагается использовать метод SVM (support vector machine, метод опорных векторов). Такой подход достаточно эффективно работает на маленьком наборе данных, в отличие от нейронных сетей, для обучения которых требуется большой обучающий корпус. Более того, за счет использования ядер, метод позволяет классифицировать даже линейно неотделимые классы путем преобразования пространства признаков, а также может быть достаточно хорошо обобщен на случай множества классов, что в рассматриваемой задаче является необходимым требованием.

Автор _____ / Быков И.А.
Научный руководитель _____ / Тимофеев Д.А.
Директор высшей школы программной инженерии _____ / Дробинцев П.Д.