УДК 004.912, 004.852

Компонент распознавания проблемно-ориентированных именованных сущностей в интеллектуально-диалоговой системе

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

В настоящее время широкое распространение получили диалоговые интерфейсы и чат-боты. При построении таких систем большую роль играет распознавание именованных сущностей двух видов: стандартные конструкции и сущности, специфичные для конкретной задачи. Например, при рассмотрении задачи планирования встреч сотрудников необходимо извлекать такие общие сущности, как имена сотрудников и время собрания, а также сущности, привязанные к конкретной области, например, название проекта, организации и т.д. Однако проблемно-ориентированные сущности встречаются в текстах заметно реже, чем общие. Проблема распознавания сущностей также возникает при решении ряда других задач, например, при автоматическом реферировании и переводе. На сегодняшний день существует несколько подходов к распознаванию именованных сущностей, которые делятся на три класса: использование методов машинного обучения, в частности, статистических и нейросетевых моделей, а также поиск сущностей в словаре и их извлечение на основе вручную разработанных правил.

На практике подходы комбинируются, однако все они требуют большого количества размеченных данных, и, соответственно, значительного объема ручной работы. Проблема усугубляется в случае языков, для которых доступно малое количество лингвистических ресурсов. Так, если для английского языка существует большое число аннотируемых и размеченных текстов, то для русского их значительно меньше. Кроме того, в диалоговых системах часто возникает задача настройки диалога для работы в новой предметной области. Для этого необходимо иметь корпусы текстов из различных предметных областей, преимущественная часть которых гораздо меньше охвачена подобными корпусами.

В качестве одного из путей решения этих проблем авторы предлагают использовать неразмеченные корпусы с последующим применением алгоритма распознавания именованных сущностей с использованием известного подхода «кластер парафраз». На первом этапе модель обучается на небольшом количестве размеченных данных, а дальше на вход такой модели подаются неразмеченные данные, в которых именованные сущности одних и тех же классов встречаются в разных контекстах. Модель объединяет такие данные в единые кластеры и для каждого определяет наиболее вероятные «триггеры» – наборы признаков, которые соответствуют конкретной именованной сущности. Если система может с достаточной степенью достоверности извлечь сущности в таком кластере, то она может получить для себя разнообразные обучающие примеры. Описанный метод был адаптирован авторами для русского языка. Найденные во время обработки неразмеченных данных сущности частично проверяются человеком. В случае неверного распознавания происходит уточнение модели.

В рамках разработки проекта интеллектуальной диалоговой системы повышения эффективности деятельности компании авторами был реализован компонент обработки текстовых сообщений пользователей для распознавания именованных сущностей в предметных областях управления задачами, управления встречами и технической поддержки сотрудников. Описанный компонент встраивается в оболочку фреймворка для конвейерной обработки естественного языка Apache UIMA в качестве аннотатора, получая на вход нормализованную реплику пользователя на естественном языке и возвращая аннотируемый файл с размеченными сущностями.

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