В. Мурашко

ПРИНЦИПЫ ОРГАНИЗАЦИИ ЗНАНИЙ В СИСТЕМЕ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ ТЕКСТОВ СМИ

Извлечение информации из текста – одна из самых востребованных задач в области автоматической обработки больших текстовых массивов. Хорошим источником для ее решения являются новостные тексты, из которых можно извлекать несколько типов данных, в частности, именованные сущности, т.е. имена людей, названия организаций, географических и геополитических объектов и т.д. Для извлечения именованных сущностей используются два основных подхода. Первый подход называется инженерным, поскольку основан на заранее созданных словарях и правилах извлечения объектов; второй – использует методы машинного обучения.

В докладе рассматриваются принципы построения лингвистической базы данных системы автоматического извлечения именованных сущностей из текстовых массивов франкоязычных СМИ на основе инженерного подхода. Материалом исследования послужили 150 текстов франкоязычных новостных сообщений, взятых с сайта *fr.euronews.com*. В ходе анализа отобранных текстов в качестве искомых объектов были определены именованные сущности таких категорий, как *персона, организация, географический объект, геополитический объект, дата, произведение искусства*, а также их всевозможные аспекты. В результате анализа были выделены конкретные лексические/грамматические маркеры и фразовые шаблоны, а также ряд правил, позволивших определить левую и правую границы именованных сущностей в текстах данного типа. Рассмотрим принципы организации перечисленных данных на примере именованной сущности персона. Так, во франкоязычном новостном сообщении при первом упоминании какой-либо персоны указываются ее полное имя и фамилия, которые выделяются заглавными буквами, например, *Jean-Luc Mélenchon, Gustavo Dudamel*. В большинстве случаев именованная сущность *персона* выполняет функцию подлежащего, реже – дополнения. Данная информация позволяет вывести правило, согласно которому в состав категории *персона* входят два и более слов, начинающихся с прописных букв и находящихся перед глаголом либо в конце предложения. Персона может обладать такими аспектами, как *должность, звание* и *профессия*, которые задаются маркерами типа *chef, président, ministre, général d'armée, colonel, aspirant, acteur, avocat*, либо шаблонами вида *le président *, le ministre *, soldat *, cycliste *, avocat **. Одним из маркеров правой границы именованной сущности категории *персона* является глагол. Если сказуемое в предложении выражено глаголом в форме прошедшего времени, то грамматическими маркерами будут формы вспомогательных глаголов *avoir* и *être* (*a, est, avait, était, n'a, n'est, n'avait, n'était*), а также их формы для местоименных глаголов (*s'est, se sont, s'était*). Отмеченные маркеры и шаблоны в полном объеме формируют часть лингвистического обеспечения системы автоматического извлечения именованных сущностей из текстов франкоязычных новостей.