Комбинирование уровней представления текста для определения признаков текстов

Куликов Василий Владимирович

Аспирант

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: segoon@cs.msu.ru

Множество задач в области автоматической обработки текстов часто решается с помощью алгоритмов машинного обучения. Эти задачи включают в себя, к примеру, определение жанра или темы текста, определение авторства или пола автора текста, определение родного языка автора текста. В случае использования машинного обучения в подобных задачах отдельные тексты преобразуются в набор признаковых векторов.

Для эффективного решения данной задачи критически важным является правильное определение признакового пространства, которому принадлежат признаковые вектора. Как правило, признаки составляются из различных свойств одного или нескольких уровней представления текста: символьного, лексического, синтаксического и т. д [1]. В таком случае в качестве признаков текстов могут выступать частоты вхождения лемм или словоформ, частоты использования символов, частоты вхождения частей речи или правил вывода грамматики языка.

Существенным недостатком разделения свойств текста на вышеуказанные уровни является неучтённость тех свойств, которые являются комбинацией свойств различных уровней, из-за чего составное свойство невозможно отнести ни к одному из уровней представления. К примеру, разделение свойств на лексический и синтаксический уровни подразумевает отдельный учёт лемм слов безотносительно их синтаксической роли и отдельный учёт правил вывода контекстно-свободной грамматики языка безотносительно лемм, участвующих в процессе вывода сентенциальных цепочек текста.

При одновременном использовании свойств слов предложения (к примеру, лемм или частей речи) и свойств дерева разбора этого предложения становится возможным учитывать и лексические свойства слов, и синтаксические конструкции языка, в которых данные слова встречаются. В частности, при использовании грамматики зависимостей для моделирования синтаксической структуры предложений русского языка исследователю доступны поддеревья дерева разбора предложения с помеченными вершинами, представляющими собой

слова и знаки препинания предложения. В таком случае для определения признаков текстов становится возможным совмещать свойства слова (лемму слова, словоформу, часть речи и т. д.) и синтаксические связи между словами с помощью этих поддеревьев, например, учитывая их частоты вхождения в текст.

При использовании корпуса текстов русских писателей XVIII–XX вв., состоящего из 47 текстов 9 авторов, и лемм пар слов, непосредственно связанных синтаксической связью, в качестве признаков текстов удалось достичь результатов скользящего контроля, превышающих показатели при использовании признаков лишь с одного уровня представления текста (униграммы и биграммы лемм, словоформ или частей речи) на 2–8% вплоть до 100%. Это свидетельствует о том, что комбинирование синтаксического и лексического уровня представления текста при определении признаков может существенно повысить точность классификации.

Литература

1. Stamatatos E., Fakotakis N., Kokkinakis G. Computer-based authorship attribution without lexical measures // Computers and the Humanities. 2001. 35. N 2. P. 193–214.
2. Романов А.С., Мещеряков Р.В. Идентификация авторства коротких текстов методами машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. “Диалог”. № 9. М.: Изд-во РГГУ, 2010. С. 407–413.