УДК 004.912

Комбинаторный метод синтаксического разбора текстов на языке эсперанто

А.Д.Чирковский
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – д. т. н, доцент, И.А. Бессмертный
(Университет ИТМО, г. Санкт-Петербург)

В настоящее время методы автоматической обработки текстов развиваются постоянно, однако это развитие характеризуется по большей части использованием английского языка в качестве целевого в связи с его распространённостью. Востребованность языка выражается в большом количестве корпусов и данных по английскому языку, что в свою очередь стимулирует разработку новых методов именно для английского языка, с учётом присущих этому языку недочётов и допущений. Эта закономерность приводит к тому, что другие в перспективе полезные языки, в частности эсперанто, оказываются незаслуженно обделены вниманием.

Так как эсперанто – язык со строгой грамматикой и морфологией без исключений из правил, он не раз представлялся как хороший инструмент в задачах обработки языка в связи с предположительной лёгкостью его обработки, например, в проекте DLT [1] вариация эсперанто использовалась в качестве промежуточного языка для машинного перевода, однако в настоящее время этот язык практически не используется, что можно также объяснить отсутствием размеченных данных для обучения статистических моделей для этого языка, кроме автоматически размеченного корпуса, доступного только через web-интерфейс [2]. Для дальнейшего же развития и возможности использования языка данные необходимы, поскольку, согласно проведённым исследованиям, несмотря на простую грамматику и хорошие показатели работы систем на правилах в таких задачах, как морфологический [3] и морфемный [4] анализ эсперанто, точность синтаксического анализа для таких систем оставляет желать лучшего (около 90%), к тому же они требовательны в плане времени и квалификации специалистов при их создании.

В этой статье рассматриваются особенности языка эсперанто, которые отличают процесс его обработки в сравнении с другими языками, обозреваются существующие исследования по обработке эсперанто и описанные в них подходы, а также предлагается новый комбинаторный метод синтаксического разбора текстов на эсперанто, учитывающий такие особенности языка, как свободный порядок слов и нерегулярную пунктуацию. Наибольшей употребимости данный метод достигает при ручной разметке текстов на эсперанто, поскольку возможно его использование в системе автоматизированной разметки для увеличения скорости работы человека, кроме того, положительными свойствами является меньшая в сравнении с традиционными системами на правилах трудоёмкость и устойчивость к проблеме синтаксического взрыва, наиболее актуальная как раз в языках с варьируемым порядком слов.

Целью данного исследования является развитие методов обработки нестандартных языков для дальнейшего улучшения качества работы систем машинного перевода на редких парах с небольшим набором данных.

Литература

1. T. Witkam. History and Heritage of the DLT (Distributed Language Translation) project. – Utrecht: Press, 2006. – 11 p.
2. E. Bick, 2007. Tagging and Parsing an Artificial Language. An Annotated Web-Corpus of Esperanto. – University of Southern Denmark: Press, 2007. – 12 p.
3. B. C. Aasgaard. Parsing of Esperanto. – University of Oslo: Press, 2007. – 144 p.
4. Theresa Guinard. An Algorithm for Morphological Segmentation of Esperanto Words. // The Prague Bulletin of Mathematical Linguistics. – 2016. – #105. – pp. 63–76.