5 Заключение

Предложен и экспериментально исследован подход к распознаванию смыслов (упоминаний целевых ситуаций, событий и фактов) в тексте, который допускает относительно простую реализацию предположительно для любого языка, при наличии возможности автоматического выделения требуемых смыслов на русском языке. Подход требует наличия корпуса квазипараллельных текстов – переводов с русского языка на иностранный или обратно. Также желательно наличие простейшего лингвистического анализатора, способного строить варианты нормальных форм для словоформ иностранного языка, что позволяет существенно повысить полноту распознавания смыслов, не требуя примеров параллельных текстов, в которых описывающие смысл слова стоят во всех возможных формах. В зависимости от видов распознаваемых смыслов от лингвистического анализатора может потребоваться умение выделять именованные сущности.

Описанные эксперименты показали высокую точность распознавания смыслов для большого количества разнообразных смыслов (40) на обучающей выборке большого объема (230 тысяч пар квазипараллельных текстов, более 1370 тысяч пар армянских и русских предложений), что, в силу особенностей выбранного способа описания смысла (*n*-ок слов, совместно встречающихся в окне), позволяет ожидать высокой точности распознавания и на других текстах. Невысокая полнота распознавания говорит о необходимости увеличить размер корпуса параллельных новостных текстов в несколько раз (с 230 тысяч пар до миллиона).

В экспериментах не использовалась контрольная выборка текстов, отличная от обучающей, для проверки полученных оценок ожидаемой точности и полноты в силу отсутствия возможности получения качественной экспертной разметки корпуса не только армянских, но и каких-либо других текстов на предмет релевантности различным смыслам. Тем не менее, просмотр содержимого построенных профилей – русских и армянских *n*-ок слов – показал релевантность подавляющего большинства из них целевым смыслам, что повышает уверенность в эффективности подхода.

Литература

[1] Eastern Armenian National Corpus, http://eanc.net

[2] Grefenstette, G. (ed.): Cross-Language Information Retrieval. Springer, 177 p. (1998)

[3] He, D., Wang, J.: Cross-Language Information Retrieval. Information Retrieval: Searching in the 21st Century, Part 11. Wiley and Sons Ltd, pp. 233-254 (2009)

[4] Nie, J-Y.: Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 3 (1), pp. 1-125 (2010)

[5] Nie, J-Y., Gao, J., Cao, G.: Translingual Mining from Text Data. Mining Text Data, Part X. Springer US, pp. 323-359 (2012)

[6] SMT Research Survey Wiki: A Comprehensive Survey of Statistical Machine Translation Research Publications. Sentence Alignment, http:// www.statmt.org/survey/Topic/SentenceAlignment

[7] Statistical Machine Translation, maintained by Philipp Koehn, http://www.statmt.org

[8] RCO Fact Extractor – инструмент компьютерного анализа текстовой информации компании «ЭР СИ О», http://www.rco.ru/?page_id=3554

[9] Ермаков, А.Е., Плешко, В.В.: Семантическая интерпретация в системах компьютерного анализа текста. Информационные технологии, (6), сс. 2-7 (2009)

[10] Ермаков, А.Е., Плешко, В.В., Митюнин, В.А.: RCO Pattern Extractor: компонент выделения особых объектов в тексте. Информатизация и информационная безопасность правоохранительных органов: Сборник трудов XII Межд. науч. конф., Москва, сс. 312-317 (2003)

[11] Потемкин, С.Б., Кедрова, Г.Е.: Выравнивание неразмеченного корпуса параллельных текстов. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Межд. конф. «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). М.: РГГУ, сс. 431-437 (2008)