Ю. И. Морозова
Yu. I. Morozova

ИЗВЛЕЧЕНИЕ ПЕРЕВОДНЫХ СООТВЕТСТВИЙ ИЗ КОРПУСА ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ДИСТРИБУТИВНОЙ СЕМАНТИКИ[1]

EXTRACTION OF TRANSLATION CORRESPONDENCES FROM A PARALLEL CORPUS USING METHODS OF DISTRIBUTIONAL SEMANTICS[1]

Аннотация. Данная работа посвящена актуальным проблемам исследования семантики лингвистических единиц с использованием корпусных методов. В работе дается описание нового направления лингвистических исследований – дистрибутивной семантики. Описывается методика извлечения переводного словаря из параллельных текстов. Предлагается расширение существующих моделей дистрибутивной семантики за счет перехода от описания лексем к описанию значимых словосочетаний.

Abstract. The paper focuses on important problems of semantic research using corpus methods. It overviews a new area of linguistic research – distributional semantics. A method for extracting a translation dictionary from parallel texts is described. The paper proposes to enhance existing models by using multi-word expressions rather than single words.

1. Обзор моделей дистрибутивной семантики

Дистрибутивная семантика – область научных исследований, занимающаяся вычислением степени семантической близости между лингвистическими единицами на основании их дистрибуционных признаков в больших массивах

[1] Данная работа выполнена при частичной поддержке гранта РФФИ 11-06-00476 «Когнитивно-лингвистические представления и разрешение неоднозначности языковых структур в системах интеллектуальной обработки знаний и машинного перевода».

лингвистических данных. Модели векторных пространств находят все более широкое применение в исследованиях, связанных с семантическими моделями естественного языка, и имеют разнообразный спектр потенциальных и действующих приложений. Основными сферами применения дистрибутивных моделей являются: разрешение лексической неоднозначности, информационный поиск, кластеризация документов, автоматическое формирование словарей (словарей семантических отношений, двуязычных словарей), создание семантических карт, моделирование перифраз, определение тематики документа, определение тональности высказывания, биоинформатика.

Теоретические основы данного направления восходят к дистрибутивной методологии З. Харриса[2,3]. Близкие идеи выдвигали основоположники структурной лингвистики Ф. де Соссюр и Л. Витгенштейн. Дистрибутивная семантика основывается на дистрибутивной гипотезе о том, что лингвистические элементы со схожей дистрибуцией имеют близкие значения[4,5].

В качестве вычислительного инструмента и способа представления моделей используется линейная алгебра. Информация о дистрибуции лингвистических единиц представляется в виде многоразрядных векторов, а семантическая близость между лингвистическими единицами вычисляется как расстояние между векторами. Многоразрядные векторы образуют

---

[2] *Harris Z. S.* Papers in Structural and Transformational Linguistics. – Dordrecht, Reidel, 1954.

[3] *Harris Z. S.* Mathematical Structures of Language. – New York, 1968.

[4] *Sahlgren M.* The Distributional Hypothesis. From context to meaning // Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), 2008, volume 20, numero 1. P. 33–53.

[5] *Turney P. D., Pantel P.* From frequency to meaning: Vector space models of semantics // Journal of Artificial Intelligence Research (JAIR), 2010, №37. P. 141–188.

матрицу, где каждый вектор соответствует лингвистической единице (слово или словосочетание), а каждое измерение вектора соответствует контексту (документ, параграф, предложение, словосочетание, слово).

Для вычисления меры близости между векторами могут использоваться различные формулы: расстояние Минковского, Манхэттенское расстояние, Евклидово расстояние, расстояние Чебышева, скалярное произведение, косинусная мера. Наиболее популярной является косинусная мера.

Существует множество разновидностей моделей дистрибутивной семантики, которые различаются по следующим параметрам:

- тип контекста (размер контекста, правый или левый контекст, ранжирование);
- количественная оценка частоты встречаемости слова в данном контексте (абсолютная частота, энтропия, совместная информация и пр.);
- метод вычисления расстояния между векторами (косинус, скалярное произведение, расстояние Минковского и пр.);
- метод уменьшения размерности матрицы (случайная проекция, сингулярное разложение и пр.).

Наиболее известными моделями дистрибутивной семантики являются латентный семантический анализ, разработанный для решения проблемы синонимии при информационном поиске[6], и модель языка как гиперпространства, разработанная как модель семантической памяти человека[7].

Концепция семантических векторных пространств (СВП) впервые была реализована в информационно-поисковой системе

---

[6] *Landauer Th. K., McNamara D. S., Dennis S., Kintsch W.* Handbook of Latent Semantic Analysis. – Mahwah New Jersey, 2007.

[7] *Lund K., Burgess C.* Producing high-dimensional semantic spaces from lexical co-occurrence // Behavior Research Methods, Instruments & Computers, 1996, 28(2). p. 203–208.

SMART[8]. Идея СВП состоит в представлении каждого документа из коллекции в виде точки в пространстве, т.е. вектора в векторном пространстве. Точки, расположенные ближе друг к другу в этом пространстве, считаются более близкими по смыслу. Пользовательский запрос рассматривается как псевдодокумент и тоже представляется как точка в этом же пространстве. Документы сортируются в порядке возрастания расстояния, т.е. в порядке уменьшения семантической близости от запроса, и в таком виде предоставляются пользователю.

Впоследствии концепция СВП была успешно применена для других семантических задач. Например, контекстное векторное пространство было использовано для оценки семантической близости слов[9]. Данная система достигла результата 92.5% на тесте по выбору наиболее подходящего синонима из стандартного теста английского языка TOEFL, в то время как средний результат при прохождении теста человеком был 64.5%.

В настоящее время ведутся активные исследования по унификации модели СВП и выработке общего подхода к различным задачам выявления семантических связей из корпусов текстов[10].

2. Извлечение переводных соответствий из параллельных текстов

Целью нашей работы является применение модели СВП для извлечения переводных соответствий из параллельных текстов.

---

[8] *Salton G.* M. The SMART Retrieval System: Experiments in Automatic Document Processing. – Prentice-Hall, 1971.

[9] *Rapp R.* Word sense discovery based on sense descriptor dissimilarity // Proceedings of the 9th MT Summit. – New Orleans, LA, 2003. – P. 315–322.

[10] *Turney P.* A uniform approach to analogies, synonyms, antonyms and associations // Proceedings of COLING, Manchester, 2008. – P. 905–912.

В работе[11] предлагается методика применения моделей дистрибутивной семантики для извлечения переводных соответствий однословных терминов из выровненных параллельных текстов. Обычно в качестве базовой информации для систем извлечения переводных соответствий используется частота совместной встречаемости терминов из соответствующих друг другу фрагментов на исходном и целевом языках. Однако предположение о том, что перевод основывается на пословных соответствиях, не соответствует действительной сложности процесса перевода. Поэтому авторы предлагают использовать в качестве минимальной единицы анализа не слово, а предложение. Лексические единицы, встречающиеся в одном предложении, связаны друг с другом синтагматическими отношениями, в то время как все предложение целиком связано с его переводом на целевой язык отношениями переводного соответствия. Поэтому каждое слово в исходном предложении связано с каждым словом в целевом предложении.

В предлагаемой модели «контекстом» для слов предложений на исходном языке выступают слова предложений на целевом языке. Контекстные векторы, описывающие слова исходного и целевого языков, помещаются в одну и ту же матрицу. Корреляция между словами вычисляется по формуле косинуса угла между их контекстными векторами. Слова из различных языков с наиболее близкими векторами считаются переводами друг друга. Данный подход особенно продуктивен, когда нужно извлечь не только самый лучший перевод данного слова, но и несколько возможных переводов.

В рамках нашего исследования был создан тестовый корпус параллельных текстов на французском и русском языках, выровненный на уровне предложений. В корпус вошли тексты научных патентов по различным темам. Объем корпуса – 100

---

[11] *Sahlgren M., Karlgren J.* Automatic Bilingual Lexicon Acquisition Using Random Indexing of Parallel Corpora // Journal of Natural Language Engineering, Special Issue on Parallel Texts, 2005, №11(3).

тысяч словоформ. Параллельные тексты были загружены в корпус-менеджер Sketch Engine[12], благодаря чему была получена морфологическая разметка текстов (леммы, части речи и грамматические характеристики).

В ходе исследования была построена модель СВП для выделения однословных переводных соответствий, которая была опробована на тестовом корпусе. Модель обладает следующими параметрами:

- тип изучаемых единиц: лексемы;
- тип контекста: предложения;
- количественная оценка частоты встречаемости изучаемой единицы в данном контексте: абсолютная частота;
- метод вычисления расстояния между векторами: косинусная мера.

Для построения векторных пространств были использованы параллельные тексты, подвергшиеся предварительной обработке:

- вместо словоформ используются соответствующие леммы;
- удалены частотные слова (в основном, служебных частей речи);
- удалены знаки препинания.

В результате применения модели СВП получен список однословных переводных соответствий, например: *moyen => средство*, *exemple => например*, *caractériser => отличать*. Приблизительная оценка точности результатов – 70%. Во многих случаях слова являются частью устойчивых словосочетаний (*par exemple => например*), такие слова не могут быть переведены правильно в рамках данной модели СВП и требуют ее усовершенствования.

3. Дальнейшие исследования

Развитие существующих подходов к построению СВП заключается в переходе к извлечению переводных соответствий

[12] **http://www.sketchengine.co.uk/**

значимых словосочетаний (ЗС) вместо отдельных лексем. В лингвистике для обозначения значимых словосочетаний используется также термин «коллокация». Мы используем данное понятие в том значении, которое принято в корпусной лингвистике, т.е. статистически устойчивые словосочетания. Для выделения значимых словосочетаний в компьютерной лингвистике используются различные статистические меры, вычисляющие силу связи между элементами в составе коллокации. Как отмечается в работе[13], мера MI (mutual information) дает наилучшие усредненные результаты. Применив меру MI к материалам тестового корпуса, мы составили частотный словарь значимых словосочетаний для предметной области научных патентов. Примеры выделенных значимых словосочетаний: *благородный металл, вспомогательное устройство, жесткий элемент, измерительная ячейка, опорный карниз, оптический луч, система охлаждения, тяжелая фракция.* В дальнейшем планируется усовершенствовать используемую модель СВП для нахождения по параллельным текстам переводных соответствий значимых словосочетаний.

4. Заключение

В работе были рассмотрены основные направления и модели нового направления исследований в компьютерной лингвистике – дистрибутивной семантики. На основании автоматической обработки больших массивов лингвистических данных возможно создание различных лингвистических ресурсов. В рамках данного направления была разработана методика извлечения однословных соответствий из параллельных текстов. Дальнейшие исследования будут связаны с применением данной методики для

---

[13] *Захаров В. П., Хохлова М. В.* Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог’2010. – М.: Изд-во РГГУ, 2010.

значимых словосочетаний, выделенных из текстов с использованием мер ассоциативной связанности слов.