Крюкова Анна Владимировна

Санкт-Петербургский государственный университет (СПбГУ),
Санкт-Петербург, Россия
krukova.ann@gmail.com
Научный руководитель – О.А. Митрофанова, канд. филол. наук

СТРОКОВЫЕ МЕТОДЫ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Ключевые слова: семантическая близость, DKPro Similarity, компьютерная лингвистика.

В работе рассматривается задача оценки семантической близости текстов на русском языке с использованием компьютерной платформы DKPro Similarity. В ходе исследования были проведены эксперименты с лексическими языконезависимыми метриками близости текстов. Результаты исследования подтверждают, что платформа DKPro Similarity пригодна для оценки семантической близости русскоязычных текстов.

Kriukova Anna

Saint Petersburg State University (SPbSU),
St. Petersburg, Russia

STRING MEASURES FOR COMPUTING SEMANTIC SIMILARITY OF TEXTS IN RUSSIAN

Keywords: semantic similarity, DKPro Similarity, computational linguistics.

The article is focused on the task of identifying the degree of semantic similarity of texts in Russian with the aid of DKPro Similarity, an open source framework for text similarity. In the course of the research experiments with string similarity measures were carried out. Results show that DKPro Similarity is applicable to computing semantic similarity of texts in Russian.

Введение

Для компьютерной лингвистики одним из важных методов, на результаты которого опирается широкий класс прикладных задач, является определение семантической близости текстов. В частности, эти данные могу использоваться при автоматической классификации и реферировании текстов, разрешении лексической неоднозначности, перефразировании и т. д. Данное направление сейчас активно разрабатывается, однако, в основном для английского языка.

Вследствие этого в нашем исследовании мы фокусируемся на оценке смысловой близости текстов именно на р у с с к о м языке и решаем эту задачу с помощью компьютерного инструмента DKPro Similarity [Bär, Zesch, Gurevych, 2013].

DKPro Similarity

Открытая и свободно распространяемая компьютерная платформа DKPro Similarity была разработана в Дармштадском Технологическом Университете (TU Darmstadt) как дополнение DKPro Core, инструмента для обработки текстов на естественном языке. В DKPro Similarity реализованы различные классы метрик близости текстов — от структурных и лексических до стилистических и даже фонетических. В данной работе используются лексические метрики (string measures), которые наименее зависят от языка входных текстов.

Используемые метрики

DKPro Similarity предоставляет реализацию более 15 строковых метрик близости, полный список которых можно найти на странице проекта на ресурсе github.com. В нашем исследовании используются семь из них, на которые авторы наиболее часто ссылаются в релевантных статьях (ср. [Mihalcea et al., 2006], [Bär et al., 2012], [Šarić et al., 2012], [Bär, Zesch, Gurevych, 2015]). Ниже приведены основные принципы их работы, более подробную информацию можно найти в литературе, приведенной выше, а также в описании метрик.

— Word N-Gram Containment Measure выражается формулой $C_n(A,B) = \frac{|Q(A,n) \cap Q(B,n)|}{|Q(A,n)|}$, где Q(A,n) и Q(B,n) — это количество n-грамм в текстах A и B (см. [Broder, 1997]);

— Word N-Gram Jaccard Measure вычисляет для двух текстов коэффициент Жаккара [Lyon, Barrett, Malcolm, 2004]: $J_n(A,B) = \frac{|Q(A,n) \cap Q(B,n)|}{|Q(A,n) \cup Q(B,n)|}$. Для метрик, использующих n-граммы, параметр n берется равным двум.

— Longest Common Subsequence Comparator находит наибольшую общую подпоследовательность с помощью определения максимального количества операций удаления или вставки символов, затем значение нормализуется [Clough, Stevenson, 2011].

— Greedy String Tiling разбивает два документа на одинаковые в них обоих не пересекающиеся друг с другом цепочки (tiles)

таким образом, чтобы ими покрывалось максимальное число символов в текстах [Wise, 1996].

— Longest Common Substring Comparator определяет наибольшую общую подстроку, опираясь на общее для текстов дерево суффиксов.
— Cosine Similarity вычисляет косинус между векторами, представляющими тексты [Manning et al., 2008].
— Levenshtein Comparator находит наименьшее число операций (вставка / удаление / замена символа), которые бы понадобились для превращения одного текста в другой [Manning et al., 2008].

Важно отметить, что все описанные метрики, кроме расстояния Левенштейна, принимают значения в промежутке [0,1], а значения Word N-Gram Containment Measure и Greedy String Tiling обусловливаются порядком, в котором тексты им передаются (поэтому для них вычисляются оба значения).

Материалы исследования

Материалами исследования выступали следующие тексты:

— аннотации статей по корпусной лингвистике из корпуса, предоставленного кафедрой математической лингвистики Санкт-Петербургского государственного университета;
— переводы романа В. Набокова «Пнин» с английского языка на русский тремя переводчиками (Б. М. Носик, С. Б. Ильин, Г. А. Барабтало);
— заголовки новостных статей из корпуса парафразов проекта «Paraphraser.ru» (каждой паре предложений, которые входят в состав этого корпуса, поставлена оценка «1», «0» или «-1» согласно тому, насколько они близки по смыслу);
— новостные сообщения из разделов «life» и «news» новостного корпуса, предоставленного кафедрой математической лингвистики Санкт-Петербургского государственного университета.

На этапе предобработки тексты были лемматизированы (с помощью Python библиотеки «PyMorphy2» [Korobov, 2015]), затем из них были удалены знаки препинания и стоп-слова. Из каждой группы случайным образом мы выбрали несколько текстов для последующей работы.

Машинное обучение

На данной стадии исследования мы рассчитали девять значений близости для всех пар документов: между собой сравнивались тексты как из одной группы, так и из разных. Следующим этапом являлась оценка полученных значений, для чего было решено использовать машинное обучение, а именно автоматическую классификацию текстов. Этот шаг разделился на следующие подзадачи:

— отделение каждого класса от остальных (то есть определение того, относятся ли два сравниваемых текста к одной группе используемых текстов или к разным);
— только для корпуса парафраз: отделение друг от друга классов предложений, относящихся отдельно к группам «1», «0», «-1»;
— только для новостного корпуса: отделение текстов, относящихся к подкорпусу «life» от относящихся к подкорпусу «news».

В экспериментах использовалось несколько линейных моделей, различия между которыми для задачи, которую решает данное исследование, не являются важными, а их общий принцип заключается в том, что модели присваивают каждому признаку (показателям близости двух текстов) некий коэффициент, и, в зависимости от значения суммы этих признаков, каждый объект (т. е. пару текстов) модель распределяет в один или другой класс. Соответственно, для каждого набора данных мы отбирали модель, результаты которой были наивысшими (оценка осуществлялась с помощью F-меры).

Результаты

Классификаторы, которые в наших экспериментах обучались на значениях лексических метрик близости, то есть достаточно простых признаках, показывают, тем не менее, неплохие результаты. Так, для нескольких наборов данных значение F-меры даже было равно единице, а именно, для тех, в которых модели различали, относятся ли к одной группе: а) соответствующие друг другу отрывки из трех переводов романа «Пнин», б) предложения из корпуса парафраз, которым поставлена оценка «1» и л и «0», и в) аннотации. Самое низкое значение, равное «0,63», оказалось у датасета, в котором классификатор отделяет от всех остальных документов сообщения из сегмента «life» новостного корпуса. Это связано с тем, что тексты в нем посвящены заметкам о небольших и незначительных событиях, а не серьезным новостям, как в подкорпусе «news», кроме

того, эти тексты касаются совершенно разных тем и при этом не используют схожие лексические конструкции. По этой причине л е к с и ч е с к и е метрики для данного датасета не могут позволить классификаторам произвести четкое разделение. У остальных наборов данных значение F-меры примерно одинаково и равно «0,86», что также является достаточно высоким показателем. Ограниченность объема данной статьи не позволяет привести здесь более подробное рассмотрение результатов исследования, однако даже приведенные сведения позволяют говорить о том, что значения лексических метрик близости являются хорошими признаками для линейных моделей при решении задачи классификации текстов.

Выводы

В данной работе мы представили инструмент DKPro Similarity, который предоставляет возможность оценивать семантическую близость текстов на русском языке. В ходе экспериментов были рассмотрены лексические языконезависимые метрики, однако эта платформа позволяет реализовать для русского языка и некоторые более сложные метрики, опирающиеся на внешние источники знаний, и мы намерены посвятить этому дальнейшую работу.

ЛИТЕРАТУРА

Bär et al., 2012 — *Bär D.* et al. UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures // SemEval '12 Proceedings of the First Joint Conference on Lexical and Computational Semantics. Vol. 1: Proceedings of the main conference and the shared task, and Vol. 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (2012). P. 435–440.

Bär, Zesch, Gurevych, 2013 — *Bär D., Zesch T., Gurevych I.* DKPro Similarity: An Open Source Framework for Text Similarity // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2013). P. 121–126.

Bär, Zesch, Gurevych, 2015 — *Bär D., Zesch T., Gurevych I.* Composing Measures for Computing Text Similarity [Technical Report] (2015). URL: http://tuprints.ulb.tu-darmstadt.de/4342/1/TUD-CS-2015–0017.pdf (дата обращения: 15.04.17).

Broder, 1997 — *Broder A. Z.* On the resemblance and containment of documents // Proceedings of the Compression and Complexity of Sequences (1997). P. 21–29.

Clough, Stevenson, 2011 — *Clough P., Stevenson M.* Special Issue on Plagiarism and Authorship Analysis // Language Resources and Evaluation. Vol. 45(1) (2011). P. 5–24.

Korobov, 2015 — *Korobov M.* Morphological Analyzer and Generator for Russian

and Ukrainian Languages // Analysis of Images, Social Networks and Texts (2015). P. 320–332.

Lyon, Barrett, Malcolm, 2004 — *Lyon C., Barrett R., Malcolm J.* A theoretical basis to the automated detection of copying // Plagiarism: Prevention, Practice and Policies Conference (2004).

Manning et al., 2008 — *Manning C. D.* et al. Introduction to Information Retrieval // Cambridge University Press, 2008.

Mihalcea et al., 2006 — *Mihalcea R.* et al. Corpus-based and Knowledge-based Measures of Text Semantic Similarity // Proceedings of the 21st national conference on Artificial intelligence. Vol. 1 (2006). P. 775–780.

Šarić et al., 2012 — *Šarić F.* et al. TakeLab: Systems for Measuring Semantic Text Similarity // SemEval '12 Proceedings of the First Joint Conference on Lexical and Computational Semantics. Vol. 1: Proceedings of the main conference and the shared task, and Vol. 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (2012). P. 441–448.

Wise, 1996 — *Wise M. J.* Yap3: Improved detection of similarities in computer programs and other texts // Proceedings of SIGCSE '96 (1996). P. 130–134.