МНОГОУРОВНЕВЫЙ ПОДХОД К ИЗМЕРЕНИЮ СТЕПЕНИ СЕМАНТИЧЕСКОГО СХОДСТВА ТЕКСТОВЫХ ФРАГМЕНТОВ

М. А. Белюга

ООО «АйЭйчЭс Глобал»
Минск, Беларусь
e-mail: marina.beliuga@ihsmarkit.com

Представлен алгоритм измерения уровня семантической эквивалентности двух текстовых фрагментов. Формирование интегральной оценки степени сходства проводилось посредством обучения регрессионной модели на основании метода опорных векторов с использованием множества признаков подобия, вычисляемых для каждой пары входных текстов.

Ключевые слова: семантическое сходство; машинное обучение; меры общности.

MULTISTAGE APPROACH FOR MEASURING SEMANTIC SIMILARITY BETWEEN TEXT SNIPPETS

M. A. Beliuha

IHS Markit Ltd
Minsk, Belarus

This paper describes the algorithm for rating the degree of semantic equivalence between two text snippets. To from the integral estimation of text similarity we use a support vector regression model with multiple features representing similarity scores calculated for each pair of sentences.

Keywords: semantic similarity; machine learning; similarity measures.

Современные исследования в области обработки естественного языка характеризуются высоким интересом к решению проблемы измерения степени смыслового подобия текстов. Решение данной задачи приобрело особую актуальность с точки зрения его применения в сферах информационного поиска, при разработке систем машинного перевода, систем автоматического аннотирования и реферирования текстов, вопросно-ответных систем, систем распознавания плагиата и др.

В свете обозначенной выше актуальности задачи многие исследователи предпринимают попытки разработки эффективных алгоритмов выявления схожих или идентичных по смыслу текстов, в том числе с использованием новейших достижений в области машинного обучения для формирования интегральной оценки смыслового подобия текстов на основании лексического сходства текстовых фрагментов без учета и с учетом порядка следования слов, их канонических форм и синонимов.

В данной работе представлен алгоритм оценки степени смыслового подобия текстов на английском языке, принципиальным отличием которого от уже существующих решений является сравнение предложений не только на лексико-грамматическом и синтаксическом уровнях, но и на уровне семантики. Алгоритм направлен на попарный анализ предложений сравниваемых текстов, с тем чтобы оценить степень их семантического сходства [1] в виде интегральной оценки по непрерывной шкале значений от 0 до 5, где оценка 0 соответствует семантически несвязанным предложениям, а оценка 5 характеризует одинаковые по смыслу пары входных предложений. Так, несмотря на лексическое и синтаксическое сходство приведенная ниже пара предложений (a) должна получать низкую оценку семантического подобия, тогда как оценка пары предложений (b) должна быть значительно более высокой вне зависимости от лексических различий:

a) «American air forces plane crashes in south Iraq» – «American air forces aircraft crashes in the south of Lybia»;

b) «Sarkozy makes official the re-election bid» – «France's Nicolas Sarkozy announces his reelection bid».

Указанная интегральная оценка степени сходства предложений формируется посредством обучения регрессионной модели на основании метода опорных векторов, реализованной в программном пакете LIBSVM[1], с использованием множества признаков подобия, вычисляемых для каждой пары входных текстов. Предполагается, что анализ семантики входных текстов способен привнести качественные улучшения в работу алгоритмов, созданных в рамках описанного выше традиционного подхода.

Понятно, что решение поставленной задачи требует формирования набора базовых лингвистических, алгоритмических и программных ресурсов. В частности, был определен базовый лингвистический процессор (ЛП) с функциональностью лексического, лексико-грамматического, синтаксического и семантического анализов текста, а также обоснована и получена лингвистическая база знаний, включающая словари и эталонные корпуса текстов из разных предметных областей (заголовки газетных статей, описания изображений, сообщения пользователей различных форумов и пр.) для разработки и анализа алгоритма решения задачи. Разработанный алгоритм был реализован с использованием лингвистического процессора IHS Goldfire [2, 3], производящего обработку текстовых документов начиная от предварительного их форматирования и заканчивая семантическим анализом. Данный лингвистический процессор ориентирован на промышленную обработку текстовых документов для целого ряда естественных языков (включая английский) и имеет лучшие на настоящее время показатели эффективности среди аналогичных разработок.

Принципиальная структурно-функциональная схема системы автоматического распознавания семантической эквивалентности текстовых фрагментов изображена на рис. 1.

[1]https://www.csie.ntu.edu.tw/~cjlin/libsvm/

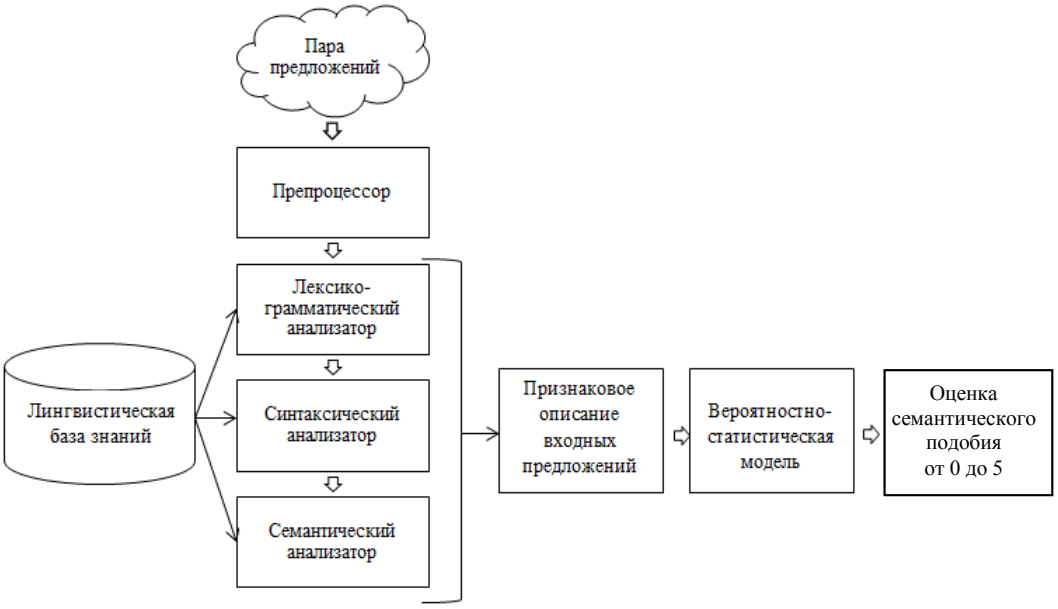

Рис. 1. Принципиальная схема решения задачи

На **предварительном этапе** при помощи ЛП IHS Goldfire текст входных предложений подвергался ряду операций нормализации:

1) поиск границ слов (токенизация);

2) нормализация написания слов и символов (*'em = them; he's = he is* и т. д.);

3) приведение различных вариантов написания кавычек к единому виду;

4) удаление стоп-слов (артикли и им подобные слова: *a, an, the, some* и т. д.; неинформативные наречия: *however, likely* и т. д.; дискурсивные маркеры: *I think, IMHO, as for me* и т. д.);

5) удаление знаков пунктуации и приведение всех слов входных предложений к нижнему регистру.

Этап лексико-грамматического анализа заключался в идентификации семантически близких и эквивалентных слов и фраз с использованием следующих ресурсов:

- базы данных лексических, фразовых и синтаксических парафраз Paraphrase Database (PPDB) [4];
- разработанные нами для решения данной задачи словари, включающие синонимичные пары для прилагательных, глаголов и существительных (*«wrong» – «incorrect»*); пары прилагательное + производное наречие (*«clear» – «clearly»*); пары действие + агенс (*«connect» – «connector»*); пары глагол + отглагольное существительное (*«pulsate» – «pulsation»*); пары существительное + образованное от него прилагательное (*«sinusoid» – «sinusoidal»*). Данные пары слов были получены автоматически с использованием деривативных аффиксов, а затем проверены на случайным образом отобранных корпусах объемом 2 млн предложений из различных областей, включая тексты Wikipedia, научных статей и диссертаций, периодических изданий и патентов;
- программного продукта Google word2vec[2].

На рис. 2 показан пример нахождения соответствующих семантически эквивалентных слов и фраз в двух предложениях с помощью вышеуказанных ресурсов.

[2]https://code.google.com/archive/p/word2vec/

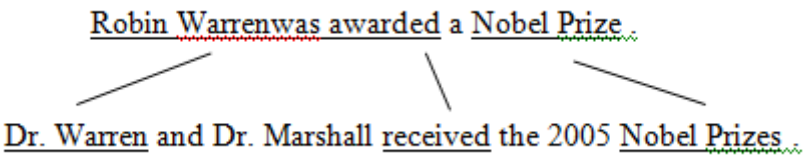

Рис. 2. Идентификация семантически близких или эквивалентных слов и фраз в паре предложений

После установления семантической корреляции между подобными словами и фразами [5] входной пары предложений проводился расчет предварительной оценки подобия предложений на основании меры Жаккара и важности слова [6] для данного предложения (TF-IDF). Мера Жаккара рассчитывалась по формуле

$$J(S_1,S_2)=\frac{|S_1\cap S_2|}{|S_1\cup S_2|}, \quad (1)$$

где S_1 и S_2 – вектора первого и второго предложений соответственно.

Мера подобия $\mathrm{sts}(S_1,S_2)$ на основании оценки TF-IDF важности слов предложений рассчитывалась по формуле

$$\mathrm{sts}(S_1,S_2)=\frac{\sum\omega\in(S_1^{\alpha}\cup S_2^{\alpha})\mathrm{tfidf}(\omega)}{\sum\omega\in(S_1\cup S_2)\mathrm{tfidf}(\omega)}, \quad (2)$$

где числитель представляет собой сумму значений TF-IDF всех слов пары предложений, для которых были установлены отношения семантического подобия, а знаменатель — сумму значений TF-IDF всех слов обоих предложений. Таким образом, повышение значимости семантически подобных слов пары предложений непосредственно влияло на повышение данной меры подобия и наоборот.

На этапе синтаксического анализа при помощи лингвистического процессора IHS Goldfire производилось построение синтаксического дерева каждого предложения из входной пары. Полученные деревья сравнивались на предмет поиска корреляции в синтаксических ролях слов, отмеченных как подобные на этапе лексического анализа [7]. Например, в паре входных строк *«the notebook of my mother»* и *«my mother cooks amazing»* синтасические роли слова *«mother»* различны: в первом случае это роль атрибута в именной группе, тогда как во втором – это роль подлежащего. Данный признак указывает на различие этих предложений на семантическом уровне.

На этапе семантического анализа вышеупомянутый лингвистический процессор использовался для выделения собственно концептов и семантических отношений между ними типа «субъект – акция – объект» (САО), где каждый элемент отношения может иметь свои атрибуты. Данные структуры полностью соответствуют таким классическим типам знаний, как объекты/классы объектов и факты об объектах. Из полученных отношений извлекались бинарные связи типа «субъект – акция», «акция – объект», «акция – непрямой объект» и другие с целью нахождения корреляции между предложениями на семантическом уровне. В данном случае на подобие предложений указывало полное или частичное совпадение «семантических ролей» слов, между которыми были установлены отношения подобия на этапе лексического анализа. При частичном совпадении, если подобные слова именных частей речи находятся в паре с глаголами, для которых отношений подобия не выявлено (например: *«predict»* и *«walk»*), то это свидетельствует о различиях на семантическом уровне. Оценка подобия предложений, формируемая на данном этапе, равнялась:

– сумме значений подобия для каждой из пар бинарных связок. Значение подобия для полностью совпадающей пары бинарных связок равнялось 1. Для частично совпадающих бинарных связок значение рассчитывалось следующим образом: 1 умножалась на долю совпадающих слов;

– числу минус 1 в случае отсутствия совпадений;

– числу 0 при отсутствии предиката и, соответственно, отсутствии отношения типа САО в одном или обоих входных предложениях.

Полученные промежуточные оценки подобия использовались в качестве признакового описания пары предложений для **обучения регрессионной модели**, необходимой для формирования интегральной оценки семантического сходства предложений в непрерывном диапазоне значений от 0 до 5.

При разработке и тестировании алгоритма использовался разработанный эталонный корпус экспертных оценок подобия для 11 306 пар предложений из разных доменных областей (заголовки публицистических статей, фрагменты научных статей, художественных текстов и др.), который был поделен на обучающую и тестовую выборки. Проведенная оценка качества работы алгоритма показала, что точность предсказаний составила 82,9 %. На наш взгляд, данный показатель свидетельствует о состоятельности выбранного подхода и позволяет надеяться на дальнейшее улучшение полученных результатов за счет более точной настройки алгоритма посредством добавления новых признаков подобия и различия предложений, необходимых для обучения регрессионной модели.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Semantic Textual Similarity, English, Spanish and Pilot on Interpretability / E. Agirre [et al.] // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015), Denver, CO, June.
2. Todhunter J., Sovpel I., Pastanohau D. System and method for automatic semantic labeling of natural language texts. U.S. Patent 8 583 422. № 12. 2013.
3. Чеусов А. В. Разработка лингвистических процессоров промышленной обработки текстовых документов // Искусственный интеллект. Интеллектуальные и многопроцессорные системы : материалы науч.-техн. конф., п. Кацивели, 25–30 сент. 2006 г.: в 3 т. / М-во образования и науки Рос. Федерации, М-во образования и науки Украины, НАН Беларуси ; ред. В. О. Бронзов. Таганрог : ТРТУ, 2006. Т. 2. С. 366–370.
4. Ganitkevitch J., Durme B. van, Callison-Burch C. PPDB: The Paraphrase 152 Database // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. P. 758–764.
5. Sultan A. md, Bethard S., Sumner T. Sentence similarity from word alignment and semantic vector composition // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval–2015). Denver, Colorado, June. Association for Computational Linguistics. P. 148–153.
6. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpeset des Jura // Bulletin de la Société Vaudoise des Sciences Naturelles 37. Lausanne. 1901. P. 547–579.
7. Oliva J., Serrano J. I., Castillo M. D. del, Iglesias Á. SyMSS: A syntax based measure for short-text semantic similarity. Data & Knowledge Engineering. 2011.