АВТОМАТИЧЕСКОЕ ВЫЯВЛЕНИЕ ОТЗЫВОВ И МНЕНИЙ В СЕТИ ИНТЕРНЕТ

М. А. Белюга

МНОГОУРОВНЕВЫЙ ПОДХОД К ИЗМЕРЕНИЮ СТЕПЕНИ СМЫСЛОВОГО ПОДОБИЯ ТЕКСТОВ

Современные исследования в области прикладной лингвистики характеризуются повышенным интересом к решению проблемы измерения степени смыслового подобия текстов. Решение данной задачи приобрело особую актуальность с точки зрения его применения в сферах цифрового и дистанционного обучения, информационного поиска, при разработке систем машинного перевода, систем автоматического аннотирования и реферирования текстов, вопросно-ответных систем, систем распознавания плагиата и многих других.

В свете обозначенной выше актуальности задачи многие исследователи предпринимали и 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, принипиальным 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 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 [1]. 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 исполь-

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

зованием 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 с функиональностью 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 проч.) для разработки и анализа алгоритма решения задачи. Так как качество ее решения напрямую зависит от точности и глубины лингвистического анализа текста, разработанный алгоритм был реализован с использованием лингвистического процессора IHS Goldfire, производящего обработку текстовых документов, начиная от предварительного их форматирования и заканчивая семантическим анализом [2; 3]. 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. Лексический, синтаксический и семантический анализ основывались на разработанной базе лингвистических правил, а также на собранных статистических данных, необходимых для получения признакового описания объектов (текстовых фрагментов), используемого в последующем для метода машинного обучения.

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

На **предварительном этапе** при помощи ЛП IHS Goldfire текст входных предложений подвергался ряду операций нормализации, упрощающих дальнейшую обработку строковой информации:

1) поиск границ слов (токенизация);
2) нормализация написания слов и символов (*don't = do not, can't = can not; 'em = them; he's = he is* и т.д.);
3) приведение различных вариантов написания кавычек к единому виду;
4) анализ содержимого скобочных контекстов внутри предложения и их представление в виде отдельных предложений в случае необходимости;
5) удаление стоп-слов:
   a. артикли и им подобные слова: *a, an, the, some, every, another, any, some, this, these, those* и т.д.;
   b. неинформативные наречия: *very, a bit, also, however, likely* и т.д.;
   c. дискурсивные маркеры: *I think, he said, IMHO, as for me* и т.д.;
6) удаление знаков пунктуации и приведение всех слов входных предложений к нижнему регистру.

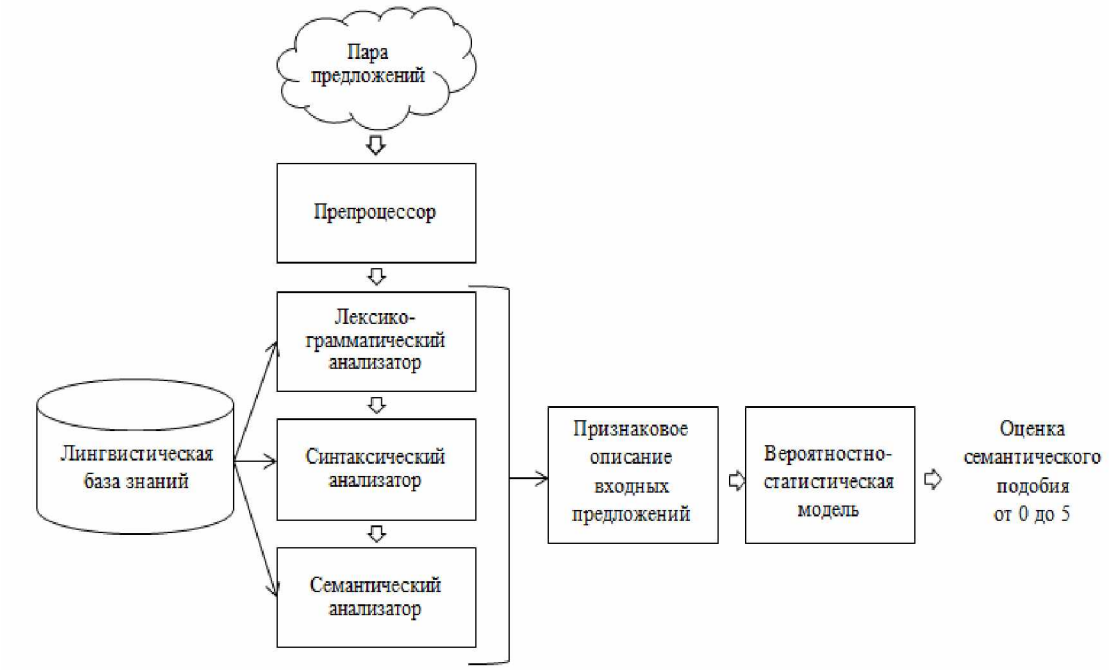

Рис. 1. Принципиальная схема решения задачи

Этап лексико-грамматического анализа заключался в идентификации семантически подобных слов и фраз при помощи:

- базы данных лексических, фразовых и синтаксических парафраз Paraphrase Database (PPDB) [4];
- разработанные нами для решения данной задачи словари, включающие:
  – 14870 синонимичных пар для прилагательных, глаголов и существительных (*"wrong" – "incorrect", "link" – "connect", "seller" – "vendor"*);
  – 1435 пар прилагательное + производное наречие (*"clear" – "clearly"*);
  – 2953 пар действие + агенс (*"connect" – "connector"*);
  – 17556 пар глагол + отглагольное существительное (*"pulsate" – "pulsation"*);
  – 563 пары существительное + образованное от него прилагательное (*"sinusoid" – "sinusoidal"*);
  – 264 пары, состоящих из названия страны и соответствующей национальности (*"uk" – "british"*);
  – 262 пары, включающих название страны и её столицу (*"uk" – "London"*).

  Деривационные списки были собраны автоматически с использованием слов из словаря ЛГК и деривативных аффиксов, а затем проверены на, случайным образом отобранных, корпусах объёмом 2 миллиона предложений из различных областей, включая тексты Wikipedia, научных статей и диссертаций, периодических изданий и патентов. При этом производное слово считалось валидным синонимом в том случае, если оно встречалось на корпусе не менее четырёх раз. Для составления синонимичных пар стран, столиц и национальностей были использованы ресурсы Wikipedia;
- программного продукта Google word2vec[1];

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

Данный этап завершался установлением семантической корреляции между подобными словами и фразами [5] с последующим расчетом предварительной оценки подобия предложений на основании меры Жаккара [6] и важности слова для каждого предложения (TF-IDF). Мера Жаккара рассчитывалась по формуле: $J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$, где S_1 и S_2 – вектора первого и второго предложений соответственно. Формирование дополнительных признаков на основании расчета меры Жаккара также производилось после этапов синтаксического и семантического анализа предложений. Мера подобия $sts(S_1, S_2)$ на основании оценки TF-IDF важности слов предложений рассчитывалась по формуле: $sts(S_1, S_2) = \frac{\sum \omega \in (S_1^{\alpha} \cup S_2^{\alpha}) tfidf(\omega)}{\sum \omega \in (S_1 \cup S_2) tfidf(\omega)}$, 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 др. с целью нахождения корреляции между предложениями на семантическом уровне. В данном случае наподобие 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 11306 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 счёт более точной настройки алгоритма посредством добавления новых признаков подобия и различия предложений, необходимых для обучения регрессионной модели.

ЛИТЕРАТУРА

1. Eneko Agirre; Carmen Banea; Claire Cardie; Daniel Cer; Mona Diab; Aitor Gonzalez-Agirre; Weiwei Guo; Inigo Lopez-Gazpio; Montse Maritxalar; Rada Mihalcea; German Rigau; Larraitz Uria; Janyce Wiebe. SemEval-2015 task 2: Semantic Textual Sim-ilarity, English, Spanish and Pilot on Interpretability. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, June.
2. *Todhunter, J.* System and method for automatic semantic labeling of natural language texts / J. Todhunter, I. Sovpel, D. Pastanohau. – U.S. Patent 8 583 422, November 12, 2013.
3. *Чеусов, А. В.* Разработка лингвистических процессоров промышленной обработки текстовых документов / А. В. Чеусов // Искусственный интеллект. Интеллектуальные и многопроцессорные системы: матер. науч.-технич. конф., п. Кацивели, 25–30 сентября 2006 г.; в 3 т.; ред. В. О. Бронзов. – Таганрог : ТРТУ, 2006. – Т. 2. – С. 366–370.
4. Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase 152Database. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 758–764.
5. Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS@CU: Sentence similarity from word alignment and semantic vector composition. In Pro-ceedings of the 9th International Workshop on Se-mantic Evaluation (SemEval 2015), pages 148–153, Denver, Colorado, June. Association for Computa-tional Linguistics.
6. *Jaccard, P.* Étude comparative de la distribution florale dans une portion des Alpes et des Jura / P. Jaccard // Bulletin de la Société Vaudoise des Sciences Naturelles 37. – Lausanne, 1901. – P. 547–579.
7. Jesús Oliva, José Ignacio Serrano, María Dolores del Castillo, Ángel Iglesias. 2011. SyMSS: A syntax-based measure for short-text semantic similarity. Da-ta & Knowledge Engineering.