УДК 004.912

АЛГОРИТМ СЕМАНТИЧЕСКОГО АНАЛИЗА ЯЗЫКОВОЙ СОСТАВЛЯЮЩЕЙ РЕЧЕВОГО СООБЩЕНИЯ

Д.В. ПЕКАРЬ, канд. техн. наук, доц. В.С. САДОВ
(Белорусский государственный университет, Минск)

Представлен алгоритм анализа языковой составляющей речевого сообщения для выявления некоторой целевой ситуации, которая заключается в выражении диктором определенного смыслового сообщения. Основой рассмотренного алгоритма является семантический граф, с помощью которого моделируется смысловой контекст выявляемой целевой ситуации. Для построения подобного графа используется лексико-семантическая база английского языка WordNet, которая позволяет осуществлять поиск связанных понятий, а также определение необходимых смысловых связей между ними. Показаны преимущества предложенного алгоритма: легкость описания целевой ситуации, отсутствие необходимости обучения алгоритму, а также учет смысловых связей в естественном языке. Результаты тестирования данного алгоритма позволяют сделать вывод о состоятельности сделанных предположений в рамках представленной работы, а также о его превосходстве перед традиционным поиском ключевых слов.

Введение. Анализ физических параметров речевого сигнала позволяет оценить эмоциональное состояние диктора, что дает возможность выявлять наступление потенциально опасных ситуаций. Для улучшения параметров работы подобных систем необходимо производить также анализ и смыслового содержания сообщений, который направлен на извлечение вербальной информации, выраженной естественным языком. Индикатором потенциальной опасности в данном случае выступает смысловое содержание сообщения, которое может иметь как прямую, так и косвенную связь с той или иной целевой потенциально опасной ситуацией. Для установления подобных смысловых связей, а также оценки их количественных параметров необходимо использовать семантический анализ, который позволяет учесть многообразие смысловых связей между отдельными понятиями естественного языка. Комбинированное использование вербальной и невербальной информации позволяет осуществлять более глубокий и комплексный анализ речевого сообщения с целью принятия более точного решения о характеристическом уровне потенциальной опасности.

Лексико-семантическая база. Основой для построения семантического анализатора сообщений служит лексико-семантическая база *WordNet* [1], созданная в Принстонском университете. Она охватывает около 155 тысяч различных лексем и словосочетаний, организованных в 117 тысяч понятий – совокупностей синонимов (synset), общее число пар «лексема – значение» составляет порядка 200 тысяч. База *WordNet* служит основой для построения других подобных лексико-семантических баз, таких как *EuroWordNet*.

Одним из центральных структурных элементов лексико-семантической базы *WordNet* является синсет (*от англ. synonym set – множество синонимов*) – множество понятий со схожим значением, служит для разграничения лексических значений и отличия их смысловых оттенков. Таким образом, понятия, входящие в синсет, образуют синонимичный ряд, который может рассматриваться как представление лексиколизованного понятия (концепта) языка.

Особенностью лексико-семантической базы WordNet является то, что понятия в ней охвачены различными связями, которые соответствуют смысловым зависимостям между концептами естественного языка. Подобное образование имеет иерархическую структуру, что позволяет формализовать понятия естественного языка. Пример такой иерархической структуры связей понятий представлен на рисунке 1.

Все лексемы языка разделены на 4 группы в зависимости от принадлежности к той или иной части речи: существительные, глаголы, прилагательные, наречия. Лексемы, принадлежащие к другим частям речи, были опущены создателями лексико-семантической базы. Лексемы различных частей речи имеют физически раздельное хранение. Такой подход объясняется тем, что лексемы различных частей речи имеют свои характерные семантические связи. В таблице приведены типы используемых семантических связей лексико-семантической базы WordNet для анализа речевых сообщений в предложенном алгоритме.

Выбор имени существительного и глагола для осуществления анализа языковой компоненты сообщения объясняется тем, что на них приходится основная смысловая нагрузка в выражении, а также названные части речи наиболее информативны для анализа.

Лексико-семантическая база WordNet позволяет осуществлять анализ понятий, входящих в выражение, а также выявлять связи с другими понятиями. Данное свойство позволяет нивелировать недостаток различных обучаемых классификаторов, а именно проблему классификации данных, отдельные признаки которых не были представлены в обучающей выборке, что часто случается при обработке языко-

вой информации из-за синонимии и других смысловых связей понятий в естественном языке. Дополнительным преимуществом использования лексико-семантической базы WordNet является отсутствие необходимости в обучающей выборке для осуществления анализа языковых выражений, что существенно облегчит использование системы на основе предложенного метода.

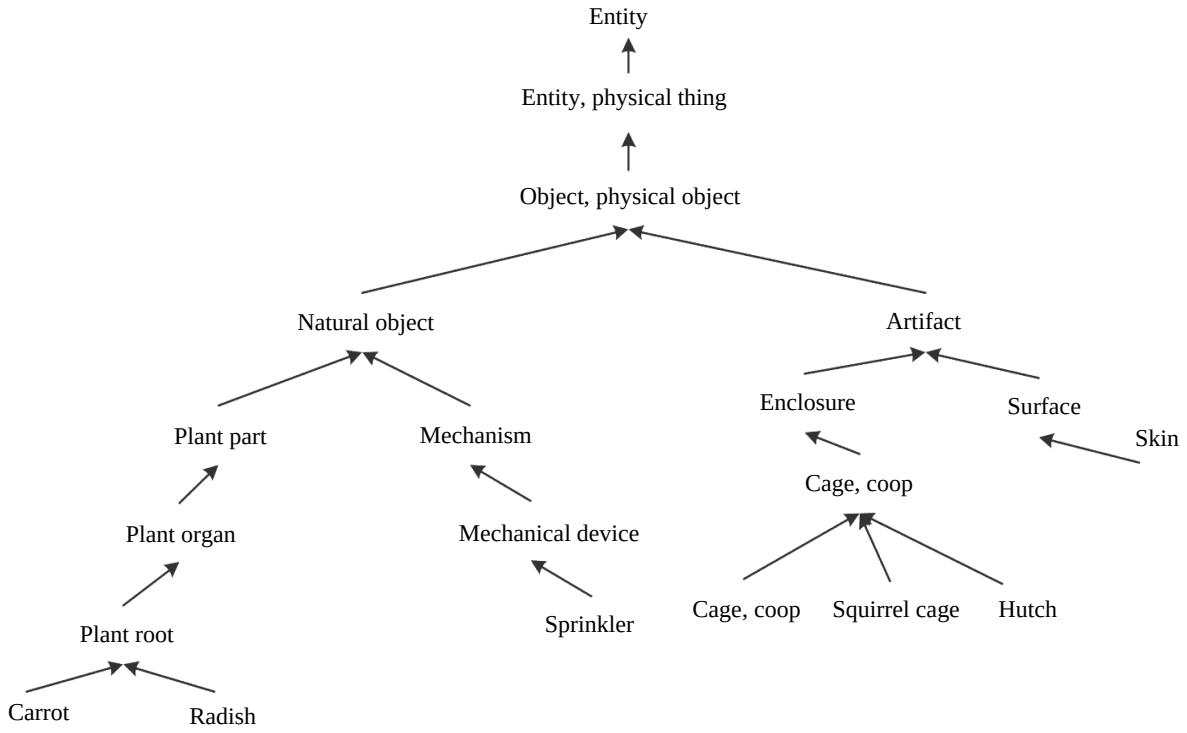

Рис. 1. Иллюстрация связей в базе WordNet

Типы используемых семантических связей в лексико-семантической базе WordNet

Часть речи	**Тип связи**	**Описание**	**Пример**
Существительное	Гипоним	Понятие, выражающее частную сущность по отношению к другому, более общему понятию	Понятие «велосипед» является гипонимом к понятию «транспортное средство»
Существительное	Гипероним	Понятие с более широким значением, выражающим общее, родовое понятие, название класса/вида предметов или понятий	Понятие «транспортное средство» является гиперонимом к понятию «велосипед»
Существительное, глагол	Синоним	Слова одной части речи, различные по звучанию и написанию, но имеющие схожее лексическое значение	Автомобиль, машина
Существительное	Мероним	Семантическое отношение понятий, при котором одно понятие выражает составную часть другого	Понятие «окно» является меронимом по отношению к «зданию»
Существительное	Холоним	Семантическое отношение понятий, при котором одно является целым по отношению к другому	Понятие «автомобиль» является холонимом по отношению к понятию «колесо»
Глагол	Логическое следование	Семантическое отношение, когда одно понятие является следствием другого	«Человек идет» – следует, что «Человек делает шаг»
Глагол	Тропонимия	Семантическое отношение, когда одно понятие является специфическим проявлением другого	Бормотать – говорить неким специальным образом
Глагол	Отношение причины	Семантическое отношение, когда одно понятие является причиной другого	Смотреть – значит видеть

Построение целевого смыслового контекста. Системы, основанные на обработке потоковой речевой информации, предъявляют специфические требования к применяемым алгоритмам анализа данных.

Условия продиктованы тем, что осуществляется анализ потоковых данных, которые не могут быть проанализированы единовременно и полностью по следующим причинам:

- информационное содержание потока формируется с течением времени;
- накопление потока не всегда реализуемо из-за возможного значительного объема данных;
- работа системы должна осуществляться в режиме реального времени, что не позволяет осуществлять анализ значительных фрагментов информационного потока;
- сложность выделения фрагмента с одним смысловым содержанием.

Для учета сформулированных причин предлагается осуществлять сегментацию входящего информационного потока на отдельные выражения или слова по мере их формирования, которые затем могут быть эффективно обработаны и проанализированы. Дополнительной аргументацией в пользу предложенного подхода является то, что анализируемый сегмент может и не содержать единого законченного смыслового посыла или содержать части логически различных или несвязанных высказываний. Анализ же отдельных понятий на предмет их соответствия или связи (а также меры близости) с некоторым целевым контекстом позволяет нивелировать отмеченное свойство.

Семантический анализ осуществляется с целью выявления предпосылок для появления той или иной ситуации, которая может быть детектирована с помощью характерных слов/понятий или упоминания какой-либо идентифицирующей информации говорящим человеком. Для формализации рассуждений полагается, что множество концептов или понятий, которые характеризуют ту или иную ситуацию, образуют смысловой контекст ситуации. Основой алгоритма является предположение, что любую ситуацию можно описать с помощью некоторого смыслового контекста, построенного путем нахождения всех понятий, семантически связанных с априорно заданными и характеризующих моделируемую ситуацию. Подобный смысловой контекст может быть представлен в виде графа, где вершинами являются множества схожих понятий одного типа, а ребра представляют собой семантические отношения между теми или иными множествами. Структура контекстного графа представлена на рисунке 2.

Рис. 2. Обобщенная структура целевого контекстного графа

Построение графа начинается с базового уровня специфичности контекста и нахождения множества синонимов для каждого из априорно заданных понятий, которые определяются пользователем для формализации целевой ситуации на естественном языке. Данный уровень принимается как базовый, поскольку в него входят исходные понятия. Вообще говоря, для каждой части речи структура семантического графа будет индивидуальной ввиду различных семантических связей для лексем различных частей

речи, однако для алгоритмизации анализа все построенные графы объединяются в один консолидированный граф, моделирующий отдельно взятую целевую ситуацию.

Согласно структуре, изображенной на рисунке 2, на базовом уровне также располагаются тропонимы и понятия, отражающие следствие некоторого действия, которые могут присутствовать только у лексем, относящихся к глаголам.

На следующих шагах происходит нахождение меронимов, гипонимов, холонимов и гиперонимов по отношению к лексемам, находящимся на базовом уровне. Очевидно, что при нахождении очередного уровня меронимов или гипонимов возрастает уровень специфичности контекста, поскольку данные уровни отражают частные понятия по отношению к базовым сущностям. При нахождении же очередного уровня холонимов или гиперонимов происходит уменьшение специфичности контекста, поскольку новые понятия отражают более общий смысл по отношению к исходным сущностям. Предложенная процедура построения семантического графа не накладывает никаких ограничений на его расширение путем добавления новых типов семантических связей между понятиями. Сформированный таким образом граф моделирует смысловой контекст той или иной ситуации.

Алгоритм семантического анализа выражений

Принимается, что исходной информацией для анализа языковой компоненты сообщения является распознанная речь в виде последовательности слов, представленных в текстовой форме. Первым этапом алгоритма является определение части речи для каждого распознанного слова из анализируемого информационного фрагмента. Данный шаг продиктован тем, что синсеты, множества гипонимов, гиперонимов, холонимов и т.д. сформированы с учетом контекста слова, а он в свою очередь зависит от части речи, к которой принадлежит употребленное слово. Для определения части речи, к которой относится то или иное анализируемое слово, применялся парсер английского языка, разработанный Стэндфордским университетом [2; 3], который обладает одними из лучших показателей точности определения и скорости обработки.

Следующим шагом на пути анализа информационного фрагмента является лемматизация – процесс определения леммы, иными словами, канонической формы лексемы. Например, слова *море*, *морем*, *моря* объединены одной общей лексемой *море*. Использование канонической формы лексемы позволяет значительно сократить число уникальных лексем (слов), а также уменьшить размер семантического графа, что оптимизирует его представление в памяти компьютера и повышает скорость работы алгоритма.

Заключительный этап анализа – проверка найденных в сообщении лемм с помощью семантического графа, моделирующего целевой смысловой контекст. Цель данного шага состоит в установлении факта наличия или отсутствия анализируемой леммы в семантическом графе. Если анализируемая лемма присутствует, происходит определение типа множества, к которому она принадлежит, а также его уровня в семантическом графе относительно уровня синонимов, тогда агрегированная оценка для всего выражения или текущего фрагмента вычисляется согласно следующему выражению:

$$S_{score} = \frac{\sum_i w_i \cdot N_i}{\sum_i N_i}, \quad (1)$$

где w_i – весовой коэффициент, который характеризует положение найденного понятия в семантическом графе; N_i – число понятий, найденных на отдельно взятой вершине семантического графа; i – индекс, отражающий связь между положением вершины графа и весом данной вершины графа при анализе речевого сообщения.

На рисунке 3 представлена обобщенная схема алгоритма семантического анализа языковой составляющей речевого сообщения.

Тестирование алгоритма

Цель тестирования алгоритма – подтверждение выдвинутого предположения о том, что целевая ситуация может быть смоделирована с помощью некоторого семантического контекста, который строится по множеству характерных понятий с использованием лексико-семантической базы. Выше было отмечено, что входными данными для семантического анализа является последовательность слов или выражений. Ввиду отсутствия значительных объемов обработанных транскрибированных речевых записей без ограничения общности были использованы текстовые базы данных записи из размеченных категорий. За ситуацию принимались упоминания соответствующей тематики в разговоре, которую необходимо выявить. Использование такого подхода давало следующие преимущества:

- категории размечены группой специалистов, что исключает влияние субъективной оценки отдельного человека;
- значительный объем данных позволяет получить робастные оценки эффективности;

- наличие достаточного количества категорий позволяет оценить устойчивость алгоритма к масштабируемости анализируемой смысловой области.

Рис. 3. Схема алгоритма семантического анализа языковой составляющей сообщения

В качестве исходных данных для тестирования использовалась текстовая база Reuters [4], которая повсеместно используется исследователями в области обработки языковой и текстовой информации. В ней было выделено 10 отдельных категорий, тексты из которых отнесены только к одной той или иной категории для исключения неопределенности оценок работы алгоритма. Для более точной имитации реальных условий работы было выбрано различное количество записей в отдельных категориях с целью недопущения появления с одинаковой вероятностью тех или иных целевых ситуаций. Структура используемого набора данных изображена на рисунке 4.

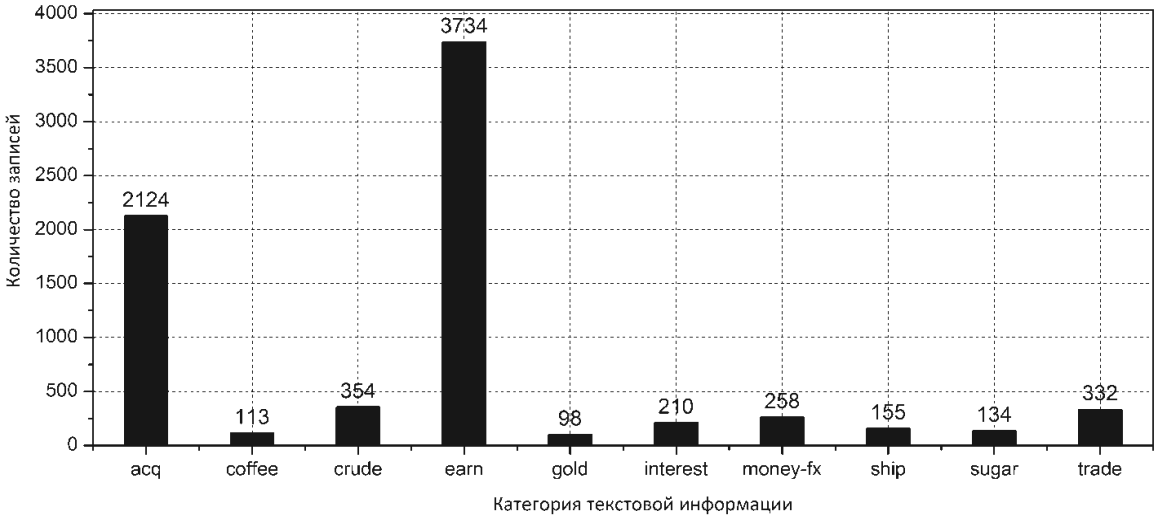

Рис. 4. Структура используемого набора данных

Как было сказано выше, для построения семантического контекста необходимы априорные понятия, которые служат основой для построения семантического графа. Поскольку выделение категорий осуществлялось без какой-то либо строгой формализации критериев, всего лишь на основе субъективной оценки группы экспертов, то в качестве априорных понятий для каждой категории принимались наибо-

лее употребляемые слова из соответствующих категорий. Для исключения варианта, когда априорно выбранные понятия могли бы служить уникальными идентификаторами той или иной категории, некоторые слова отнесены одновременно к нескольким категориям. Таким образом, исключается возможность однозначного распознавания контекста по «понятию-ключу». Количество общих понятий, вычисленное согласно выражению (2), составило 49,3 %:

$$P_{shared} = \frac{N_{shared}}{N_{total}} \cdot 100\ \%, \tag{2}$$

где N_{shared} – число понятий, встречаемых в нескольких категориях; N_{total} – общее количество используемых понятий.

Следующим этапом является построение целевого контекста для каждой категории согласно описанному выше алгоритму. Построение семантического графа осуществлялось до 1-го уровня.

На первом этапе исследования проводилась оценка эффективности обработки языковой информации без учета семантических связей между отдельными понятиями. Для осуществления данной цели выполнялась кластеризация записей текстовой базы данных, т.е. группировка записей на основе сходства употребления тех или иных понятий в каждой тестируемой записи. После формирования кластеров, а их количество задавалось априорно и равнялось числу используемых категорий, проверялось, верно ли запись была отнесена к тому или иному кластеру по априорно заданной категории. Анализ осуществлялся с помощью программного инструментария *Weka* [5] с использованием методов *KMean* [6] и *EM* [7] кластеризации. Результаты анализа приведены на рисунке 5.

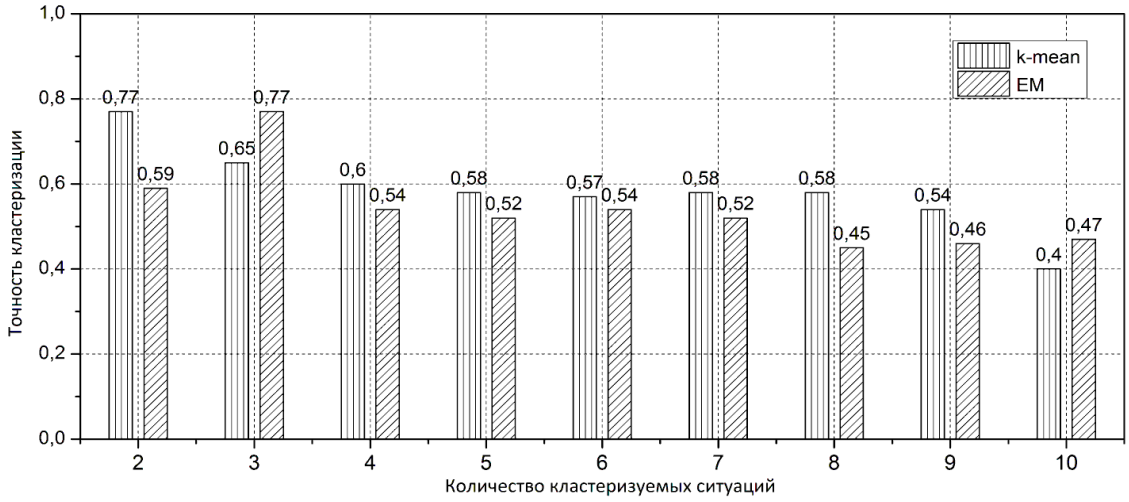

Рис. 5. Результаты точности кластеризации тестовых записей

На основе полученных результатов можно сделать вывод о том, что определение схожести сообщений только на основе сравнения частоты использования ключевых слов не дает высокой точности обработки. Полученный результат подтверждает тот факт, что выражение одной и той же мысли может осуществляться различными словами и понятиями, связанными различными семантическими отношениями между собой, которые не учитываются статистическим методом – нахождением характерных выражений на основе максимальной частоты их употребления.

Следующий этап тестирования – нахождение точности определения категории сообщений путем сравнения схожести их содержания с априорно заданным контекстом из семантически связанных понятий, а в реальном применении – способности выявления целевой ситуации с использованием семантического графа. Для моделирования реальных условий работы тестирование осуществлялось путем анализа в произвольном порядке записей текстовой базы данных. Значение точности определения категории рассчитывалось согласно следующему выражению:

$$P_{precision} = \frac{N_{correct}}{N_{total}}, \tag{3}$$

где $N_{correct}$ – число правильно распознанных ситуаций; N_{total} – общее число записей в текстовой базе данных.

В приведенном тестировании N_{total} = 7512.

Результаты тестирования приведены на рисунке 6.

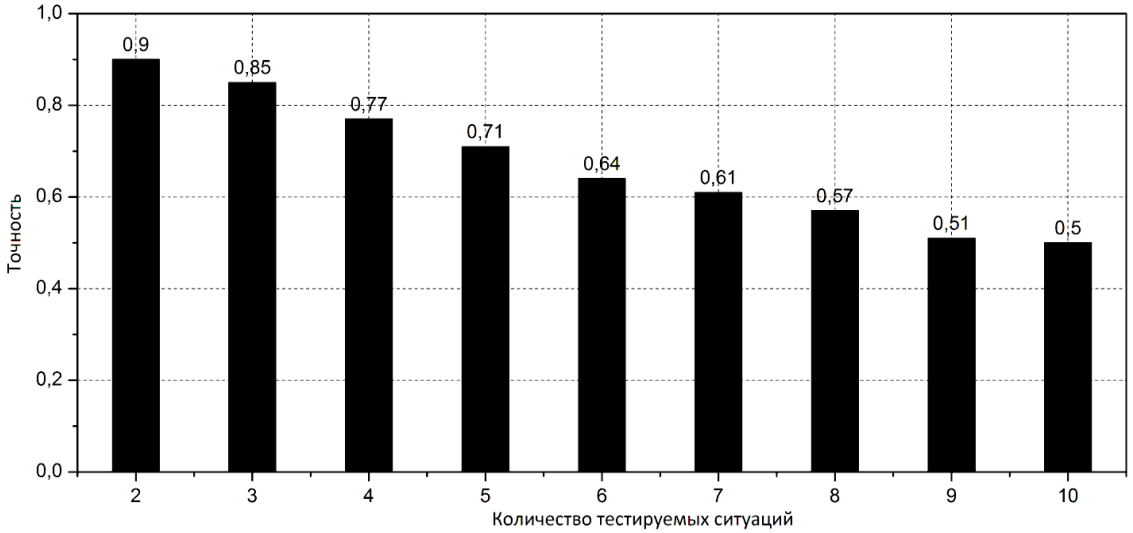

Рис. 6. Результаты тестирования предложенного алгоритма

Заключение. Разработанный алгоритм семантического анализа языковой составляющей сообщения показал превосходство перед традиционными алгоритмами обработки языковой информации. Важным преимуществом предложенного алгоритма является его применение в реальных задачах без этапа обучения. Положительный результат тестирования обоснован успешной практической проверкой состоятельности предположения о возможности моделирования целевой ситуации с помощью множества семантически связанных понятий.

ЛИТЕРАТУРА

1. Introduction to WordNet: An On-line Lexical Database [Electronic resource]. – Режим доступа: http://wordnetcode.princeton.edu/5papers.pdf. – Дата доступа: 26.09.2012.
2. Toutanova, K. Enriching the Knowledge Sources Used in a Maximum Entropy: Part-of-Speech Tagger / K.Toutanova, Ch.D. Manning // In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. USA. – 2000. – P. 63 – 70.
3. Toutanova, K. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network: In Proceedings of HLT-NAACL / K. Toutanova, D. Klein, Ch. Manning. – USA. – 2003. – P. 252 – 259.
4. Machine Learning and Intelligent Systems, University of California [Electronic source]. – Access mode: http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection. – Date: 06.02.2011.
5. Holmes, G. WEKA: a machine learning / G. Holmes // Intelligent Information Systems. – 1994. – № 29. – P. 357 – 361.
6. Jain, A.K. Data Clustering: 50 Years Beyond K-Means / A.K. Jain // Pattern Recognition Letters. – 2010. – № 8. – Vol. 31. – P. 651 – 666.
7. Clustering With EM and K-Means [Electronic resource]. – Режим доступа: http://cseweb.ucsd.edu/~atsmith/project1_253.pdf. – Дата доступа: 25.10.2012.

Поступила 15.01.2013

SEMANTIC BASED ALGORITHM FOR LANGUAGE COMPONENT ANALYSIS OF SPEECH

D. PEKAR, V. SADOV

The presented study describes an algorithm for analysis of the language component of speech to detect a certain situation which is connected with some expressed meaning. The basis of the proposed algorithm is a semantic graph which helps to model the semantic context of the detected situation. To construct this graph lexical-semantic database of English WordNet is used which allows to search for related concepts, and to identify the necessary semantic relations between them. The proposed approach has several advantages like the easiness of describing and modelling of target situations, there is no need to train the system, and also taking into account semantic relations within natural language. The obtained test results allow to say that the suggestions made are justifiable and the proposed algorithm outperforms the traditional key-word based approach.