

ISSN 1561-5359

ИСКУССТВЕННЫЙ Т НАУЧНО-ТЕОРЕТИЧЕСКИЙ ЖУРНАЛ ИНТЕЛЛЕКТ



2 ' 2007

ISSN 1561-5359

Національна академія наук України
Інститут проблем штучного інтелекту

**ШТУЧНИЙ
ІНТЕЛЕКТ**

2'2007

Национальная академия наук Украины
Институт проблем искусственного интеллекта

**ИСКУССТВЕННЫЙ
ИНТЕЛЕКТ**

2'2007

National Academy of Sciences of Ukraine
Institute of Artificial Intelligence

**ARTIFICIAL
INTELLIGENCE**

2'2007



ІІШІ МОН і НАН України «Наука і освіта»

Засновники журналу **Національна академія наук України,**
Інститут проблем штучного інтелекту НАНУ і МОНУ

Головний редактор

Анатолій Іванович Шевченко,
чл.-кор. НАН України, професор, доктор технічних наук, доктор богослов'я,
директор Інституту проблем штучного інтелекту

Редакційна колегія

Л.А. Білозерський, к.т.н.
С.М. Вороний, к.т.н.
В.П. Гладун, професор, д.т.н.
Ю.І. Журавльов, академік РАН (Москва)
І.А. Каляєв, професор, д.т.н. (Таганрог)
Ю.В. Капітонова, професор, д.ф.-м.н.
І.М. Коваленко, академік НАНУ
Ю.В. Крак, д.ф.-м.н., професор
Роман Куц, професор, Єльський університет (Нью-Гейвен, США)
С.В. Машенко, к.т.н.
К.М. Нюнькін, к.ф.-м.н.
Ю.І. Самойленко, чл.-кор. НАНУ
В.І. Скурихін, академік НАНУ
В.М. Ткаченко, с.н.с., д.т.н.
В.І. Черній, професор, д.мед.н.
А.О. Чикрій, чл.-кор. НАНУ
В.Ю. Шелєпов, д.ф.-м.н.
А.П. Шпак, академік НАНУ

Відповідальний редактор

С.Б. Іванова, заступник директора
Інституту проблем штучного інтелекту

Відповідальний секретар

І.С. Сальников, кандидат технічних наук,
вчений секретар Інституту проблем штучного інтелекту

Свідоцтво про державну реєстрацію КВ № 1803 від 20.11.1995 р., ISSN 1561-5359
Журнал «Штучний інтелект» внесено до переліку журналів ВАК України, у яких можуть публікуватися результати дисертаційних робіт на здобуття наукових ступенів кандидата й доктора наук за спеціальностями «Фізико-математичні науки» та «Технічні науки»

*Рекомендовано до друку вченою радою Інституту проблем штучного інтелекту
МОН і НАН України. Протокол № 5 від 28 травня 2007 р.
Електронна версія попередніх номерів знаходиться на web-сервері інституту (м. Донецьк)
<http://www.iai.donetsk.ua>*

УДК 004.89:004.4

А.А. Егошина

Донецкий государственный институт искусственного интеллекта, Украина

Об одном способе построения статического словаря морфологического процессора

Предложены состав и структура статического словаря морфологического процессора полнотекстовых информационно-поисковых систем с естественноречевым интерфейсом. Для ускорения поиска часто используемые основы и соответствующие аффиксы помещаются в кэш-словарь.

В 60 – 70-е гг. XX века центральным звеном экспериментальных исследований в области машинного морфологического анализа являлось создание машинного словаря. Не существовало единого общепринятого формата и структуры такого словаря. В результате этого обстоятельства каждый алгоритм разрабатывался под определенный формат словаря и являлся словарнозависимым. В современных интеллектуальных информационно-поисковых системах морфологический процессор стал неотъемлемой частью.

Основная проблема при разработке лексического и алгоритмического обеспечения компонентов морфологического процессора состоит в хранении словаря большого объема и соответственно большого времени поиска в нем. Исследования в этой области направлены на минимизацию исходных данных. Работы, посвященные морфологии, можно условно разделить на две категории:

- 1) теоретические – в некоторых из них представлены описания морфологических законов и формальные модели русской морфологии;
- 2) прикладные – дано описание программно-реализованных систем с морфологическим модулем.

Теоретические работы посвящены построению многоуровневых формальных моделей морфологии и в большинстве своем предназначены для синтеза. Подобные модели морфологического синтеза подразумевают наличие больших словарей со сложной структурой. Они описывают широкий круг морфологических явлений (фонетическая реализация слова, акцентная парадигма, большое число словообразовательных аффиксов). Недостатком таких моделей является их сложность: несколько уровней представления морфологической информации, специальные грамматики для перехода с одного уровня на другой, избыточность грамматических признаков, часть из которых выделена в модели для описания частных случаев [1].

Модели, которые используют словарь, способны дать более полный анализ словоформы (т.е. оперировать большим числом грамматических признаков). Степень точности такого анализа выше по сравнению с моделями, которые не используют словарь. На пространстве реальных текстов системы, использующие словарь, могут часто давать сбои. Это обусловлено тем, что не существует полных словарей. Лексика языка непрерывно пополняется – появляются новые слова. Для каждой предметной области существует своя терминология, свое подмножество лексики языка, и включить в общий словарь всю существующую терминологию – невозможно.

Существует два базовых подхода к проектированию морфологических машинных словарей (лексиконов) для флективных языков.

Первый копирует академическую лингвистическую модель описания, где выделяются основные парадигматические классы, соответствующие типу склонения и спряжения, и правила регулярных альтернатив (фонетических чередований), а нерегулярные формы (например, сильные глаголы в немецком и английском языках) задаются перечислением. Такого типа лексиконы для русского языка составляются на базе модели грамматического словаря А. Зализняка, разрабатывая 8 классов именного склонения и 16 глагольного спряжения, а чередования в основе и глагольной темы выносятся в отдельное множество постморфологических правил альтернатив.

Второй подход рассматривает любого вида регулярное и нерегулярное чередование как часть расширенной псевдофлексии (в таком случае, основа словоформы 'день' – 'д', а флексия – '-ень'; для словоформы 'песок': 'пес' и '-ок'). В подобной модели описания число парадигматических классов для русского языка возрастает до 3000, но рост числа классов при проектировании компенсируется однородностью лексикона и отсутствием как исключений, так и правил альтернатив.

Внутреннее устройство лексиконов первого и второго типов не влияет ни на процесс лемматизации – приведения словоформы к нормальной форме слова, репрезентирующей лексему – ни на морфологический анализ. Анализаторы, построенные на разных типах лексиконов, могут одинаково эффективно использоваться как для морфологического анализа, так и для синтеза [1].

Постановка задачи. Целью работы является разработка состава и структуры статического словаря морфологического процессора для полнотекстовых ИПС с естественноречевым интерфейсом, а также определение возможности использования полученной модели для более полного и точного анализа словоформ. Целью и результатом морфологического анализа является определение морфологических характеристик слова и его основной словоформы. Перечень всех морфологических характеристик слов и допустимых значений каждой из них зависит от естественного языка. Тем не менее, ряд характеристик (например, название части речи) присутствует во многих языках. Результаты морфологического анализа слова неоднозначны, что можно проследить на множестве примеров.

В полнотекстовых ИПС с естественноречевым интерфейсом обрабатываются запросы двух классов: запросы добавления новых документов и запросы пользователя на поиск документов в существующей базе документов. При обработке запросов каждого класса основными функциями, реализуемыми процессором морфоанализа, являются: получение всех словоформ слова, постановка слова в заданную форму и получение грамматических характеристик словоформы.

Для реализации этих функций морфологический процессор содержит основные модули, показанные на рис. 1.

Основными компонентами процессора являются:

- модуль разделения текста на составляющие;
- модуль анализа словоформ;
- статический морфологический и кэш-словари.

Модуль разделения текста на составляющие

Принимает исходный текст документа или текст от компонентов пользовательского интерфейса. Анализируемое предложение попадает на вход модуля разделения текста в виде массива символов, содержащего прописные и строчные буквы

русского алфавита, цифры, знаки пунктуации. Полученный массив преобразуется в массив лексических единиц. Для каждой лексической единицы формируется отдельная строка, в которую копируются все символы, принадлежащие данной лексической единице. При этом удаляются пробелы, символы переноса, конца строки и незнакомые символы. В зависимости от результатов обработки полученная цепочка символов направляется в один из трех потоков данных:

- цифровые или символьные сокращения ('см', '15.06.06');
- аббревиатуры ('ДГИИИ');
- полные словоформы.

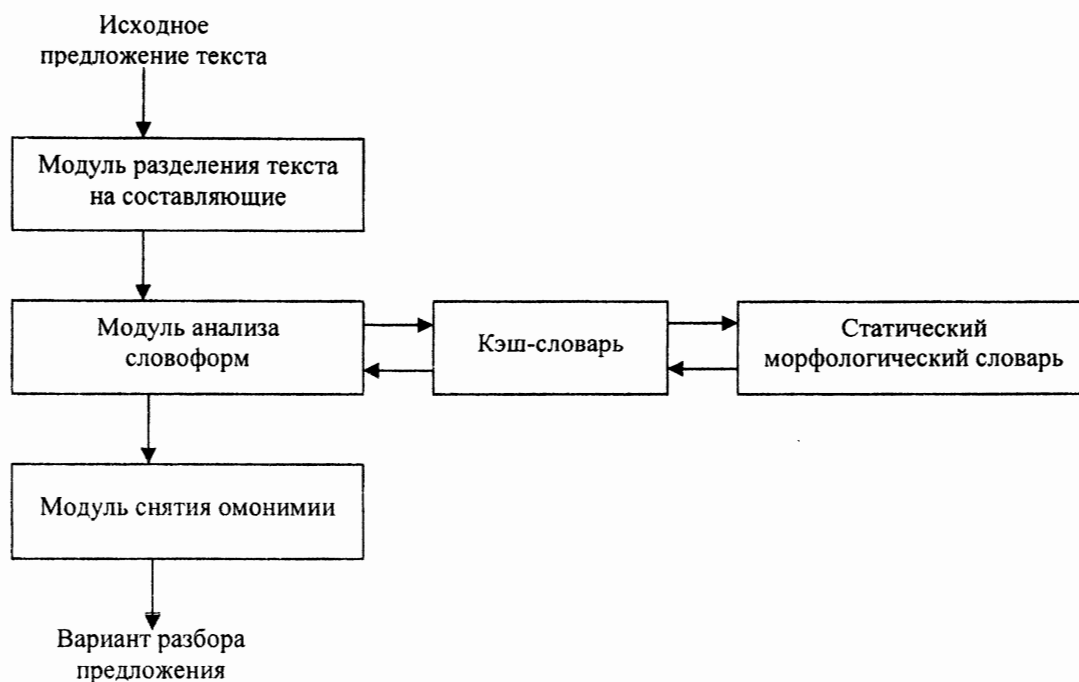


Рисунок 1 – Компоненты морфологического процессора

Первые два потока данных считаются проиндексированными и не подвергаются дальнейшему морфологическому анализу.

Полные словоформы поступают на вход модуля анализа словоформ, цель которого – разбить все множество словоформ на подмножества по признаку принадлежности к той или иной лексеме (множеству словоформ, отличающихся друг от друга только словоизменительными значениями [2]), привести все элементы каждого такого подмножества к уникальной основе, однозначно определить грамматические характеристики лексемы и проиндексировать тексты по встретившимся в них основам.

Модуль анализа словоформ

На вход поступает массив лексических единиц, выделенных из входного текста модулем разделения. Для каждой лексической единицы выполняется поиск аффиксов в статическом морфологическом словаре, состав которого представлен на рис. 2.

При этом вначале в наборе суффиксов подбирается соответствующий исходному слову, запоминается его вид (например, суффикс существительного). Затем из слова выделяется предшествующая суффиксу часть. Это может быть сочетание корня и приставки (префикса) или только корень. На следующем шаге выполняется аналогичный поиск-проверка на наличие приставки, если она есть – отделяется. На последнем этапе выполняется поиск корня, т.е. оставшаяся после разбиения часть слова ищется в наборе корней.



Рисунок 2 – Состав статического морфологического словаря

Если в результате поиска не найдено ни одного успешного варианта, то проводится поиск среди исключений, также хранящихся в словаре. Разница между исключениями и обычными основами состоит в том, что словообразование исключений происходит нестандартным образом. В связи с этим в словаре хранятся не морфемы исключений, а слова целиком. Таким образом, при поиске среди исключений приходится просматривать весь набор слов. Это занимает много времени, поэтому поиск среди исключений проводится только в том случае, когда не найдено ни одного варианта среди обычных основ.

В случае, когда все этапы поиска дали отрицательный результат (не найдено ни одного варианта), пользователю выдается запрос на ввод новой основы в словарь. В случае его отказа это сделать выполнение морфологического анализа прекращается. Если же новое слово введено в словарь, то вся процедура поиска повторяется [3].

В результате работы модуля анализа словоформ получаем набор морфем, из которых состоит исходное слово, и его принадлежность к определенной части речи.

Статический морфологический и кэш-словари

Прообразом морфологического словаря является словарь А.Н. Тихонова [4]. Он содержит информацию об основах, аффиксах и исключениях. Все выделяющиеся в слове части имеют ту или иную семантику: непроеводная основа выражает основное лексическое значение слова, а остальные морфемы (их называют служебными морфемами, или аффиксными) – дополнительное лексическое и грамматическое значение. Все множество морфем русского языка делится по разным основаниям на несколько классов. В классификации учитывается следующее: роль морфемы в слове, значение морфем, их место в слове, их происхождение. Выделяются корневые морфемы и аффиксальные. Основой для такого членения есть место и роль

таких морфем в слове. Корневые морфемы – это обязательная часть слова. Без корня не существует слов. Аффиксальные морфемы – это факультативная часть слова. Аффиксы, входя в слово, относят его к какой-нибудь разновидности, к какому-нибудь классу предметов, признаков, процессов. В этом и заключается принципиальное различие между аффиксальными и корневыми морфемами – обязательная повторяемость аффиксов в аналогично построенных и обладающих общим элементом значения словах и безразличие к этому свойству корней [5]. Семантический принцип организации подобного словаря позволяет использовать при анализе значений слов основной принцип объектно-ориентированного программирования – наследование.

Для ускорения поиска часто используемые основы и соответствующие аффиксы помещаются в кэш-словарь.

Метод морфологического разбора словоформы позволяет значительно уменьшить требования к памяти, так как требует хранения только основных частей словоформ и таблицы флексивных частей (префиксов, суффиксов и инфиксов). Требования к оперативной памяти уменьшаются в 50 – 200 раз, но из-за использования более сложных алгоритмов сопоставления увеличиваются требования к производительности процессора.

Литература

1. Еськова Н.А., Бидер И.Г. Формальная модель русской морфологии. – Ч. 1-2. – М., 1978.
2. Егошина А.А. Языковые и алгоритмические аспекты построения морфологических процессоров для интеллектуального поиска в полнотекстовых базах данных // Труды IV Междунар. конф. «Интеллектуальный анализ информации ИАИ-2006». – К.: Просвіта. – 2006. – 334 с.
3. Смирнов Ю.М., Андреев А.М., Березкин Д.В., Брик А.В. Об одном способе построения синтаксического анализатора текстов на естественном языке // Изв. вузов. Приборостроение. – 1997. – Т. 40, № 5. – С. 34-42.
4. Тихонов А.Н. Словообразовательный словарь русского языка. – М.: Русский язык, 1985.
5. Лефевр В.А., Земская Е.А. Современный русский язык и словообразование. – М.: Просвещение, 1973.
6. Мельчук И. Курс общей морфологии. – М., 1997. – Т. 1.

Г.А. Егошина

Про один спосіб побудови статичного словника морфологічного процесора

Подана розробка складу та структури статичного словника морфологічного процесора повнотекстових інформаційно-пошукових систем з природномовним інтерфейсом. Для прискорення пошуку часто основи, що використовуються, та належні афікси розміщуються у кеш-словнику.

A.A. Yegoshina

On the Mode of Construction Morphological Process Static-line Dictionary

The composition and structure syntax dictionary of morphological full-text information storage and retrieval system with natural language interface are proposed. To accelerate the search some often used fundamentals and corresponding affixes are included in cache memory.

Статья поступила в редакцию 11.07.2006.

Авторы номера

Агарков А.В.	Институт проблем искусственного интеллекта МОН и НАН Украины, г. Донецк, aav@iai.donetsk.ua
Аль-Аммори Али, к.т.н.	Национальный транспортный университет, г. Киев, Украина
Бабаков Р.М., к.т.н., доцент	Донецкий государственный институт искусственного интеллекта, Украина, cpld@mail.ru
Баркалов А.А., д.т.н., профессор	Университет Зеленогурский, Польша, a.barkalov@iie.uz.zgora.pl
Бармак О.В.	Хмельницький національний університет, Україна
Берднік П.Г.	Харківський університет Повітряних Сил імені Івана Кожедуба, Україна, bpgma@list.ru
Берко А.Ю., к.т.н., доцент	НУ «Львівська політехніка», Україна, BerkoAndriy@Yandex.lv, victana@bk.ru
Висоцька В.А.	НУ «Львівська політехніка», Україна
Грамотина О.В.	Донецкий государственный институт искусственного интеллекта, Украина
Двухглазов Д.Е.	Харківський університет Повітряних Сил імені Івана Кожедуба, Україна, bpgma@list.ru
Егошина А.А.	Донецкий государственный институт искусственного интеллекта, Украина
Єфімов Г.М.	Таврійська державна агротехнічна академія, м. Мелітополь, Україна
Жук А.В.	Институт проблем искусственного интеллекта МОН и НАН Украины, г. Донецк
Иванова С.Б.	Институт проблем искусственного интеллекта МОН и НАН Украины, г. Донецк, sbi@iai.donetsk.ua
Качур І.В., к.б.н.	Донецький державний інститут штучного інтелекту, Україна
Кравец Т.Н., к.ф.-м.н., доцент	Донецкий государственный институт искусственного интеллекта, Украина
Крак Ю.В., д.ф.-м.н., профессор	Київський національний університет імені Тараса Шевченка, Україна
Курилов М.А.	Институт проблем искусственного интеллекта МОН и НАН Украины, г. Донецк, kurilov ukr@mail.ru
Кхальфауи Нежиб бен Хамади	Донецкий национальный технический университет, Украина
Межейникова Л.С.	Институт проблем искусственного интеллекта МОН и НАН Украины, г. Донецк, megeinikova@mon.gov.ua
Мурыгин К.В., к.т.н.	Институт проблем искусственного интеллекта МОН и НАН Украины, г. Донецк, kir@iai.donetsk.ua
Нищенко А.В.	Институт проблем искусственного интеллекта МОН и НАН Украины, г. Донецк
Ньюнкин К.М., к.ф.-м.н.	Институт проблем искусственного интеллекта МОН и НАН Украины, г. Донецк, kmn@iai.donetsk.ua
Острова Т.В., к.мед.н.	Донецький державний медичний університет на ім. М. Горького ДОКТМО, Україна Донецький державний інститут штучного інтелекту, Україна
Павленко М.А.	Харківський університет Повітряних Сил імені Івана Кожедуба, Україна, bpgma@list.ru
Пряничникова Е.А.	Донецкий государственный институт искусственного интеллекта, Украина
Резников В.А., к.т.н., доцент	Донецкий государственный институт искусственного интеллекта, Украина
Рішняк І.В.	НУ «Львівська політехніка», Україна
Руденко В.М.	Харківський університет Повітряних Сил імені Івана Кожедуба, Україна, bpgma@list.ru
Сарычев А.П., к.т.н.	Институт технической механики НАН Украины и НКА Украины, г. Днепропетровск, Saryuchev@prognoz.dp.ua
Селяков Е.Б.	Донецкий государственный институт искусственного интеллекта, Украина
Ткаченко М.Д.	Харківський університет Повітряних Сил імені Івана Кожедуба, м. Харків, Україна
Цыганаш В.Е., к.т.н., доцент	Донбасская государственная машиностроительная академия, г. Краматорск, Украина

Содержание

Научное событие.....	3
Раздел 1	
Моделирование объектов и процессов	
<i>Баркалов А.А., Бабаков Р.М., Кхальфауи Нежиб Бен Хамади</i>	
Моделирование кэш-памяти микрокоманд с прямым отображением данных	6
<i>Берко А.Ю., Рішняк І.В., Висоцька В.А.</i> Моделі оцінювання ризиків інформаційної безпеки в системах електронної комерції.....	14
<i>Крак Ю.В., Бармак О.В., Єфімов Г.М.</i> Синтез мимічних виразів емоцій на основі формальної моделі	22
<i>Павленко М.А., Руденко В.М., Двухглазов Д.Е. Берднік П.Г., Ткаченко М.Д.</i> Метод побудови інтелектуальних моделей імітації поведінки повітряних об'єктів	32
<i>Резников В.А., Пряничникова Е.А.</i> Об экспертном управлении плохо формализуемыми объектами	40
<i>Сарычев А.П.</i> Моделирование в классе систем регрессионных уравнений со случайными коэффициентами в условиях структурной неопределенности. Часть 1. Случай независимых коэффициентов	48
Раздел 2	
Системы и методы искусственного интеллекта	
<i>Азарков А.В.</i> Поиск изоморфных пересечений двух графов за полиномиальное время.....	62
<i>Аль-Аммори Али</i> Методика обеспечения эффективности информационно-управляющих систем сигнализации опасных полетных ситуаций.....	75
<i>Егошина А.А.</i> Об одном способе построения статического словаря морфологического процессора	87
<i>Иванова С.Б., Межейникова Л.С., Курилов М.А.</i> Дидактические аспекты реализации электронных изданий учебного назначения	92
<i>Кравец Т.Н., Грамотина О.В.</i> Расчет нетто-премий и резервов в негосударственном медицинском страховании.....	99
<i>Мурыгин К.В.</i> Обнаружение объектов на изображении на основе каскада классификаторов	104
<i>Нюнькин К.М.</i> Поиск автомобильного номерного знака на изображении	109
<i>Селяков Е.Б.</i> Функциональные особенности мобильного измерительно-управляющего комплекса	119
<i>Цыганаш В.Е.</i> Особенности решения задачи нормализации изображений с использованием концепции интеллектуальной системы	124
<i>Черній В.І., Острова Т.В., Качур І.В.</i> Застосування методу нейросітьового моделювання для класифікації електроенцефалограми.....	130
<i>Шелепов В.Ю., Ниценко А.В., Жук А.В.</i> Новые алгоритмы распознавания фонем и их классов, поиск слова по его смешанной транскрипции при распознавании слов большого словаря	139
АВТОРЫ НОМЕРА.....	148
АВТОРСКИЙ УКАЗАТЕЛЬ.....	149
РЕЦЕНЗЕНТЫ НОМЕРА	149
ДО ВІДОМА АВТОРІВ	151

**ЧИТАЧУ!
ПЕРЕДПЛАТИТЬ НАШ ЖУРНАЛ!**

- ◆ якщо Вас цікавлять теоретичні проблеми штучного інтелекту,
- ◆ якщо Ви хочете знати про розробки українських спеціалістів і вчених СНД у цій новій галузі,
- ◆ якщо Ви здатні захопитися складними технічними завданнями та їх творчим вирішенням

Передплатний індекс УДПНЗ «Укрпошта» 23563

Передплатний індекс НВН «Ідея» 10118

Журнал «Штучний інтелект» виходить 4 рази на рік

**ЧИТАТЕЛЬ!
ПОДПИШИТЕСЬ НА НАШ ЖУРНАЛ!**

- ◆ если Вас интересуют теоретические проблемы искусственного интеллекта,
- ◆ если Вы хотите знать о разработках украинских специалистов и ученых СНГ в этой новой области,
- ◆ если Вы способны увлечься сложными техническими задачами и их творческим решением

Подписной индекс УТНПС «Укрпочта» 23563

Подписной индекс ННП «Идея» 10118

Журнал «Искусственный интеллект» выходит 4 раза в год

У журналі публікуються статті українською, російською та англійською мовами

Технічний редактор В.М. Пігуз

Коректори К.С. Івашко, Л.О. Глущенко, О.М. Трубнікова, О.В. Сукова

Комп'ютерна верстка Н.О. Захарова, Н.В. Лащенко, О.В. Петюренко, Г.П. Чепурина, П.П. Волянський

Здано до набору 04.05.2007. Підписано до друку 06.07.2007. Формат 70×108/16.
Обл.-вид. арк. 14,82. Наклад 300 прим. Зам. № 75/07/120 від 04.05.2007

Оригінал-макет виготовлено в редакційно-видавничому відділі
Інституту проблем штучного інтелекту МОН і НАН України
Україна, 83050, м. Донецьк, пр. Б. Хмельницького, 84

Інститут проблем штучного інтелекту,

e-mail: edoffice@iai.donetsk.ua

<http://www.iai.donetsk.ua>

Віддруковано в Інституті проблем штучного інтелекту (Свідоцтво про внесення суб'єкта видавничої справи до державного реєстру видавців, виготівників і розповсюджувачів видавничої продукції № 444, серія ДК від 08.05.2001 р., вид діяльності у видавничій справі – видавнича діяльність, виготовлення видавничої продукції, розповсюдження видавничої продукції).

**Національна академія наук України
Інститут проблем штучного інтелекту**