УДК 004.853, 004.55

О.Н. Лесько, Ю.В. Рогушина

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ ДЛЯ СНЯТИЯ ОМОНИМИИ В ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТАХ

Разработан метод снятия различных видов омонимии в естественно-языковых текстах деловых, научных и правовых документов. Специфика метода заключается в том, что он не требует использования большого количества синтаксических правил и корпуса размеченных текстов, что значительно упрощает его реализацию и сокращает время, необходимое для создания и разметки корпуса текстов. Этот результат достигается как за счет использования онтологии предметной области, так и за счет особенностей синтаксических структур деловых, научных и правовых документов.

Ключевые слова: омонимия, морфологический анализ, синтаксический анализ, обработка естественно-языковых текстов, онтологии.

Введение

На современном этапе развития информационных систем большую роль приобретает использование в них знаний о *предметных областях* (ПрО). Это вызывает потребность в создании средств формализации таких знаний и методов их извлечения из разнообразных *информационных ресурсов* (ИР).

Поскольку основным источником знаний о ПрО являются *естественно-языковые* (ЕЯ) документы, то именно автоматизация их обработки для извлечения содержащихся в них знаний является одним из приоритетных направлений в таких исследованиях.

Сложность такой обработки обуславливается неоднозначностью естественного языка. В частности, одной из проблем, возникающих при распознавании семантики ЕЯ-текста, является необходимость обнаружения и разрешения омонимии. В данной работе рассматриваются различные виды омонимии и способы ее разрешения, которые используют знания о ПрО, представленные в виде онтологии, и ограниченный набор правил.

Проблема омонимии в естественно-языковых текстах

Одной из проблем, возникающей при анализе ЕЯ-текстов, являются слова-омонимы. Омо́*нимы* (др.-греч. ὁμός – одинаковый + ὄνομα – имя) – разные по значению, но одинаковые по звучанию и написанию слова. В лингвистике различаются несколько видов омонимов, но в данной работе мы будем рассматривать два из них – омоформы и омоморфемы.

Омоформы – слова, совпадающие только в отдельных формах. Это могут быть слова как одной части речи (например, «засипати» (от «спати») и «засипати» (от «сипати»)), так и разных («шию» (от «шити») – «шию» (от «шия»)). *Омоморфемы* – части слов (приставки, суффиксы, корни, окончания), совпадающие в написании и произношении, но имеющие разные значения (например, окончание «и» в украинском языке может означать как именительный падеж множественного числа, так и родительный падеж единственного числа – «річк*и*»). Омонимы, отличающиеся только одной или двумя формами, в данной статье не рассматриваются.

В конструкциях, содержащих синтаксические омонимы, реализуется двойная (множественная) связь, при которой одно слово или группа слов подчиняется любой из доминант, но при этом изменяется семантика высказывания. Хотя такое явление встречается довольно часто как в русском, так и в украинском языке, исследований, направленных на изучение синтаксической неоднозначности, до 60-х го-

дов прошлого столетия в отечественном языкознании не было [1].

Следует отметить, что при всей сложности синтаксиса естественного языка в документах официально-делового стиля и в обычных информационных текстах используется относительно небольшое количество синтаксических структур, например, несогласованное определение может встретиться в виде причастного/деепричастного оборота или придаточного предложения, но не группы существительного. Кроме того, в таких текстах используется только прямой порядок слов, и практически не встречаются непроективные синтаксические конструкции [2].

Методы снятия омонимии

Как указано, например, на сегодняшний день существуют системы разрешения неоднозначности, основанные на правилах, и вероятностные системы [3, 4]. Системы, основанные на правилах, развиваются с 60-х годов прошлого века, выполняют локальный или глобальный синтаксический разбор. Вероятностные системы [5], использующие статистику совместной встречаемости грамматических признаков слов в больших корпусах, омонимия в которых снята заранее.

Вероятностные системы требуют наличия общедоступного корпуса текстов, которого для украинского языка нет. Получение большого объёма морфологически размеченных текстов вручную – задача крайне трудоёмкая, поэтому обычно для разметки текстов используют заранее сконструированные морфологические анализаторы (например, Mystem, Pymorphy). Однако автоматические разметчики, как правило, приписывают слову не единственный разбор, а все теоретически возможные. Для английского языка, как для языка с бедной морфологией, задача снятия морфологической омонимии сводится, как правило, к проблеме разрешения многозначности на уровне частей речи (так называемого POS-теггинга). При этом используются алгоритмы, основанные на статистических моделях, учитывающие вероятность появления тега той или иной части речи в данном контексте. Для английского языка эти алгоритмы работают достаточно хорошо и обычно демонстрируют не менее 96 % точности, ошибаясь лишь в 4 % случаев [6].

Для русского и украинского языков точность таких алгоритмов намного меньше. Во-первых, морфологическая омонимия в русском языке, не сводится к омонимии частей речи, а охватывает множество различных грамматических признаков. Во-вторых, в английском языке порядок слов фиксированный.

Это позволяет, к примеру, опираться только на локальный контекст слова (соседние слова) без учета дальних зависимостей. Поэтому для морфологической дизамбигуации в английском языке можно успешно использовать алгоритмы, основанные на Марковских моделях и учитывающие зависимость каждого набора тегов только от одного элемента контекста – непосредственно предшествующего ему набора тегов. В русском языке количество возможных контекстов из-за этого увеличивается и эффективность обучения простой модели, основанной на локальных зависимостях, снижается. Поэтому, наряду с Марковскими моделями, для снятия морфологической омонимии в русском языке используются более сложные статистические модели или гибридные системы, в которых статистика дополняется набором правил [6]. Здесь также приводятся результаты сравнения использования скрытой Марковской модели (HMM) и Марковской модели максимальной энтропии (MEMM) для решения проблемы морфологической дизамбигуации в русском языке. Результативность снятия морфологической омонимии в русском языке примерно на том же уровне, что и при работе с английским материалом, т. е. около 90 %. При этом точность немного меняется в зависимости от того, какой набор тегов частей речи подаётся алгоритму на вход. С задачей дизамбигуации по расширенному набору тегов, напротив, оба алгоритма справляются не очень хорошо, не превышая порога точности в 90 %.

Системы, базирующиеся на правилах, требуют описания большого количества правил, и поэтому весьма сложны для реализации и работы в реальном времени. Практически все существующие алгоритмы снятия омонимии на основе правил включаются в состав синтаксического анализа, что создает трудноразрешимое противоречие, когда для успешного снятия омонимии необходимы точные результаты синтаксического анализа, для получения которых, в свою очередь, нужно предварительно снять омонимию. Кроме того, значительный объем исходного числа связей существенно замедляет обработку, приводя к так называемому «комбинаторному взрыву». При таком подходе строятся все возможные варианты синтаксического разбора, что приводит к увеличению времени разбора (поскольку одна словоформа может иметь множество вариантов морфологических характеристик). Одна из таких систем описана в [7]. Кроме того, приведенный здесь алгоритм является не последовательным, а параллельным, и для его реализации недостаточно средств обычных языков программирования.

Для снятия синтаксической омонимии возможно использование семантических знаний, либо использование знаний, полученных на основе совместной встречаемости слов в корпусе текстов. В работе [8] описана система, в которой для однозначного определения синтаксической структуры входного предложения используется лексико-семантическая онтологическая база знаний UkrRusWordNet.

Для методов, основанных на синтаксических правилах, характерны следующие недостатки сложность формального описания этих правил, особенно для флективных языков со свободным порядком слов, и снятие омонимии выполняется на этапе синтаксического анализа, что означает повторение процедуры синтаксического анализа для каждого из вариантов омонимичной словоформы.

Недостатками вероятностных методов являются длительность формирования и разметки корпуса текстов и невысокая точность анализа, вызванная свободным порядком слов во флективных языках.

Таким образом возникает необходимость в создании метода снятия неоднозначности для омоформ разных частей речи и омоморфем-окончаний, не требующий ни большого числа правил, ни корпуса размеченных вручную текстов.

Поэтому возникает необходимость в разработке гибридного метода, использующего как правила, так и информации из текстов, опубликованных в интернете, не требующего повторений процедуры синтаксического анализ в случае наличия омонимии. Также используется семантическая информация (одушевленность, информация о том, что слово является именем человека или названием организации).

Постановка задачи

Цель данного исследования – это снятие семантических неоднозначностей в ЕЯ-текстах, в частности, связанных с омонимией. Для этого необходимо разработать алгоритм синтаксического анализа текста, позволяющий однозначно определить морфологические характеристики определенного слова в предложении.

Поскольку существующие алгоритмы сложны для практической реализации и использования, для разработки такого алгоритма предлагается использовать не только морфологическую информацию, но и онтологию ПрО.

Особенностью разрабатываемого алгоритма является то, что он не требует ни большого корпуса текстов, как вероятностные методы, ни большого числа правил, как формальные методы, что значительно расширяет сферу его использования.

Использование онтологий для автоматической обработки текстов

Для формального представления знаний отдельных ПрО сегодня широко используются онтологии. Онтологию можно рассматривать как *базу знаний* (БЗ) специального вида с семантической

информацией об определенной ПрО [9]. Компоненты, из которых складываются конкретные онтологии, зависят от парадигмы представления, но практически все модели онтологий содержат определенные концепты (понятие, классы), свойства концептов (атрибуты, роли), отношение между концептами (зависимости, функции) и ограничения использования, которые определяются аксиомами. Формальная модель онтологии О представляет собой тройку O=<T,R,F>, где Т – множество понятий ПрО; R – множество отношений между ними; F – множество функций интерпретации понятий и отношений. Такая модель может быть конкретизирована в зависимости от назначения и сферы применения онтологии.

Фундаментальные понятия определенной ПрО должны соответствовать *классам* онтологии. Для определения *экземпляра* достаточно объявить его членом какого-либо класса.

При обработке информации на естественном языке (ЕЯ) часто используются специализированные онтологии. Основное их назначение в таких задачах – обеспечить связь между фрагментами текста на ЕЯ и понятиями ПрО (например, классами или экземплярами онтологии). В частности, широко используются тезаурусы и лингвистические онтологии [10].

Особенности подмножества синтаксиса естественного языка для деловых и законодательных документов

Минимальной коммуникативной единицей в ЕЯ-тексте является предложение. В качестве коммуникативной единицы предложение описывает факт действительности. Каждое предложение состоит из одного или нескольких слов, объединенных в соответствии с законами грамматики и характеризуется грамматическим единством, а также относительной смысловой и интонационной завершенностью. Основным признаком предложения является предикативность, т. е. в предложении должен быть предикат (Р) и субъект (S). Предикат может означать процесс, действие, состояние, признак, свойство, которым обладает субъект. Участник, на которого распространяется действие или на которого направлено отношение, является объектом (Obj). Эти, а также другие семантические отношения между предикатом и другими элементами высказывания составляют модель управления предиката, и в естественных языках выражаются с помощью особого вида синтаксических связей – управления.

В украинском языке, как и в других флективных языках, синтаксическое управление реализуется в виде флексий (окончаний). Субъект в текстах деловых документов выражается существительным в именительном падеже, предикат выражается глаголом или тире (в случае обозначения видовой принадлежности). Переходные глаголы требуют наличия существительного в винительном падеже, которое обозначает объект действия [11].

Грамматическое единство и смысловая завершенность позволяют людям анализировать предложения, в которых некоторые отдельно взятые слова и флексии могут интерпретироваться неоднозначно. Но в общем случае составление набора формальных правил для автоматического снятия таких неоднозначностей довольно сложно, поскольку множество с синтаксических структур естественного языка бесконечно ([2]). Но тексты законодательных документов обладают особенностями, которые позволяют составить относительно небольшой набор правил для их анализа. Приведем некоторые наборы из этих особенностей: 1) по интонационному оформлению предложения в законодательных документах могут быть только повествовательными; 2) в таких текстах применяется только прямой порядок слов; использование несогласованных определений ограниченно и сводится в основном к причастным и деепричастным оборотам; 3) глубина вложенности таких определений ограничена в основном первым уровнем. Вместе с требованиями синтаксического единства и грамматической завершенности в эти особенности позволяют составить набор правил,

для снятия морфологической омонимии в официально-деловых текстах.

Эти правила такие:

– для непереходных глаголов в личной форме в предложении должно быть существительное в именительном падеже и не должно быть винительного падежа;

– для переходных глаголов в личной форме в предложении должно быть по одному существительному в именительном и винительном падеже;

– в предложении (причастном или деепричастном обороте) может быть только по одному существительному в дательном, творительном и предложном падежах;

– для безличных глаголов и предикатов в предложении не должно быть существительных в именительном падеже;

– причастные, деепричастные обороты и однородные члены предложения выделяются запятыми;

– могут использоваться контекстные правила, например, согласование прилагательного (указательного местоимения) и следующего за ним существительного в роде, числе и падеже.

Для снятия морфологической омонимии может использоваться онтология ПрО. Например, если в онтологии ПрО есть термин «землі сільськогосподарського призначення», при этом слово «землі» употреблено в именительном падеже, везде в тексте, где встречается это словосочетание, слову «землі» соответствует именительный падеж, но не родительный или винительный.

Для снятия синтаксической омонимии возможно использование семантической информации, например, онтологии, либо совместную встречаемость словоформ в документе, либо корпусе текстов. Например, в предложении «дохід з джерелом їх походження з України – будь-який дохід, у тому числі, але не виключно, доходи у вигляді доходів страховиків – резидентів від страхування ризиків страхувальників – резидентів за межами України» именная группа «за межами України» может относиться к именной группе «страхування ризиків» или «страхувальників – резидентів». Но, поскольку в документе встречается словосочетание «страхування ризиків за межами України», можно сделать соответствующий выбор.

Алгоритм анализа предложения

Рассмотрим алгоритм анализа предложения в общем виде. Для анализа возвратных глаголов, предложений с частицей «нет» и предикатов причастий существует несколько особых правил.

Алгоритм анализа предложения в случае отсутствия омонимии состоит из следующих шагов:

– морфологический анализ предложения. Результатом является массив (однозначных) морфологических характеристик каждого слова в предложении;

– поиск глагола, причастий, деепричастий и отглагольных существительных. Результатом является список глаголов;

– определение модели управления для каждого из найденных на предыдущем этапе слов.

Модель управления для переходных глаголов состоит из шести падежей, для непереходных – из пяти падежей (кроме винительного). При этом именительный либо именительный и винительный являются обязательными. Для причастных (деепричастных) оборотов именительный падеж является недопустимым. Результатом является список обязательных и возможных падежей для каждого глагола.

Далее для каждого из найденных на предыдущем этапе слов выполняется следующая процедура:

– начиная от выделенного глагола, причастия, деепричастия или отглагольного существительного просматриваются все слова слева направо до тех пор, пока морфологическая информация слова содержится в модели управления (т. е. пока какой-то из падежей не встретился повторно). Результатом является причастный (деепричастный) оборот либо именная группа с отглагольным существительным;

– проверяется наличие обязательных падежей. Результатом является двоичное значение (истина/ложь).

Алгоритм анализа предложения при наличии омонимии содержит следующие шаги:

– морфологический анализ предложения. Результатом является массив (однозначных) морфологических характеристик каждого слова в предложении;

– поиск глагола, причастий, деепричастий и отглагольных существительных. Результатом является список глаголов;

– определение модели управления для каждого из найденных на предыдущем этапе слов. Модель управления для переходных глаголов состоит из шести падежей, для непереходных – из пяти падежей (кроме винительного). При этом именительный либо именительный и винительный являются обязательными. Результатом является список обязательных и возможных падежей для каждого глагола;

– далее для каждого из найденных на предыдущем этапе слов выполняется следующая процедура:

- начиная от выделенного глагола, причастия, деепричастия или отглагольного существительного просматриваются все слова слева направо до тех пор, пока морфологическая информация слова содержится в модели управления (то есть пока какой-то из падежей не встретился повторно). Результатом является причастный (деепричастный) оборот либо именная группа с отглагольным существительным;
- составление системы уравнений для выделенной глагольной группы (причастного или деепричастного оборота);
- решение системы уравнений для выделенной глагольной группы.

Процедура составления системы уравнений для выделенной глагольной группы заключается в следующих действиях:

– выполняется просмотр каждого слова w_n. Если морфологическая информация слова неоднозначна и перед словом есть предлог, проверяется его сочетаемость с глаголом. Для этого используется таблица сочетаемости предлогов и падежей;

– $X_{n_i} = 0$, если падеж i не содержится в морфологической информации слова n;

– $X_{n_i} = 1$, если падеж i содержится в морфологической информации слова n, где $i = \overline{1,6}$ и соответствует падежам.

Результатом является система уравнений, поскольку в выделенном сегменте может быть только по одному слову каждого падежа, и каждому слову должен соответствовать только один падеж.

Предлагается следующая процедура решения системы уравнений.

Пусть V_r – вектор обязательных падежей. Может быть {1, 0} или {1, 0} в зависимости от переходности.

Пусть V_0 – вектор возможных падежей,

$V_{0_i} = 0$ или $V_{0_i} = 1$, $i = \overline{1,6}$.

Если вектор X состоит из одного элемента (в предложении только одно существительное), ему необходимо присвоить обязательный падеж $X = X \cap V_r$. Если $\sum X_i = 1$, анализ закончен и предложение однозначно. Иначе предложение неоднозначно.

Иначе:

для каждого уравнения $X = X \cap V_r$.

находим j, для которого $\sum X_{i_j} = 1$.

1. Для каждого X находим $\mathrm{X} = \mathrm{X} \cap \mathrm{X_j}$.

2. Если для всех X $\sum \mathrm{X_i} = 1$, то процесс анализа закончен и предложение однозначно.

3. Иначе переходим к 1.

Особенности предложенного метода

Специфика предложенного в работе метода связана со сферой его использования.

1. Алгоритм предназначен в основном для анализа научных текстов и текстов документов. Такие тексты обычно

содержат только простые предложения, сложные предложения, состоящие из двух простых, а также причастные и деепричастные обороты. Кроме того, такие тексты обычно не содержат явных синтаксических ошибок, например, в падежных формах, что упрощает процесс анализа, и непроективных синтаксических структур.

2. Использование семантической информации, например, о том, что слово является именем человека или названием организации, переходным или непереходным глаголом. Такая информация находится в словаре системы.

3. Использование информации о сочетании глаголов и предлогов. С этой целью используется поиск сочетаний глаголов и предлогов в Google.

4. Не строится полная синтаксическая структура предложения, и не рассматриваются все возможные варианты морфологических характеристик слов.

5. Снятие омонимии происходит в после морфологический анализ до выполнения синтаксического анализа, а не на этапе синтаксического анализа.

Архитектура модуля снятия омонимии в поисковой системе «Правотекст»

Основными модулями системы являются: морфологический словарь, модуль морфологического анализа, модуль распознавания поименованных сущностей, онтология ПрО, таблица сочетаемости глаголов предлогов с падежами, модуль сегментации предложения, модуль составления системы уравнений и модуль решения системы уравнений. Также в системе используются словарь словоформ, содержащий словоформу и морфологическую информацию (укр.), словарь словоформ и исходных форм, который содержит информацию о канонических формах слов.

Для морфологического анализа текстов на украинском языке используется собственная база данных (морфологический словарь), созданная на основе материалов сайта “http://lcorp.ulif.org.ua/dictua/”.

Для снятия омонимии существительных с предлогами используется таблица сочетаемости глаголов предлогов с падежами. Таблица построена по толковому словарю [12].

Модуль распознавания поименованных сущностей [13] находит имена и названия в тексте и обозначаемые ими сущности.

Онтология ПрО, использующая для выделения их терминов в тексте.

Модуль сегментации предложения, выделяющий причастные и деепричастные обороты, а также однородные члены предложения.

Модуль составления системы уравнений – составляет систему уравнений по результатам морфологического анализа.

Модуль решения системы уравнений – находит единственный правильный вариант морфологической информации.

Вначале работы алгоритма исходное предложение обрабатывается модулем морфологического анализа. Результатом обработки является список исходных форм каждого слова и морфологическая информация.

Затем выполняется поиск терминов ПрО (однословных и многословных) в онтологии ПрО. Результатом является список терминов в предложении и морфологическая информация для них. Например, в онтологии есть термин «оскарження рішень контролюючих органів». Если в тексте найдено это словосочетание, слову «рішень» всегда приписывается родительный падеж.

Морфологическая информация каждого слова передается модулю сегментации предложения. Результатом является набор глагольных групп (причастных и деепричастных оборотов и т. д).

Затем модуль составления системы уравнений обрабатывает каждую глагольную группу. Результатом является набор уравнений

$$\sum X_{i_j} = 1,$$

где i – номер слова в предложении, j – номер морфологической информации. Затем эта система уравнений решается соответствующим модулем.

Пример результатов работы алгоритма для предложений на украинском языке показан на рисунке.

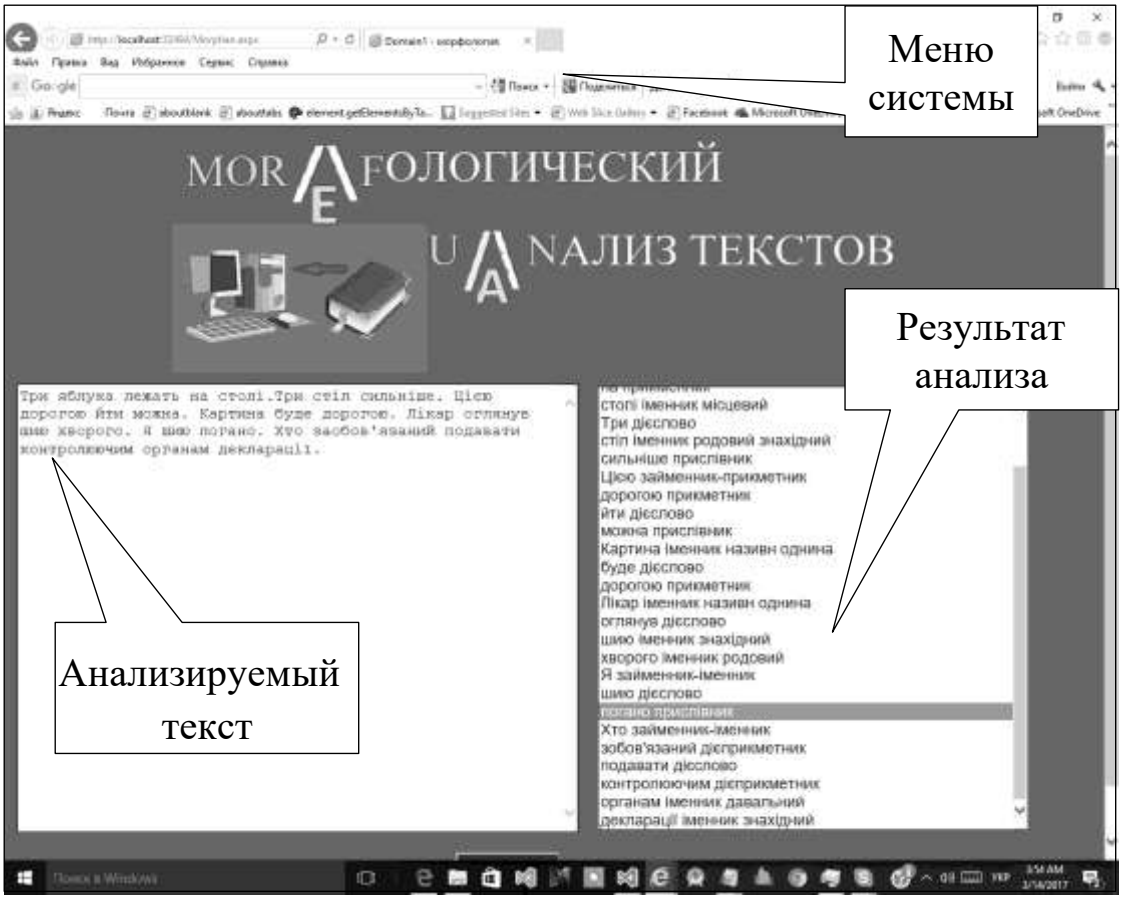

Рисунок. Результат работы алгоритма

Перспективы использования разработанного алгоритма

Одна из важных задач, находящихся за рамками классического управления знаниями, – это сопоставление онтологий с природными текстами. Ее подзадачами являются разметка ЕЯ текста терминами онтологии, пополнение онтологии знаниями, добытыми из размеченного ЕЯ-текста, и вычисления степени семантической близости между текстом и онтологией. Решение этой задачи требует учета специфики отдельных естественных языков, и поэтому существующие средства и методы решения этой задачи должны разрабатываться для каждого языка отдельно. Решение этой задачи может быть одной из областей применения данного алгоритма.

Построение лингвистической БЗ, которая позволяет соотносить фрагменты ЕЯ-текста с терминами онтологии, также находится, но не рассматривается в данной работе (следует отметить, что на сегодня существует определенное количество таких БЗ, в том числе и для украинского языка, и средств их использования в семантической разметке). Для решения этой задачи тоже может использоваться данный алгоритм.

Данный алгоритм используется в модуле нормализации терминологии системы автоматизированного построения онтологии [13]. Напомним, что в этой работе описана система автоматизированного построения онтологии на основе текста Налогового Кодекса Украины. Для построения онтологии необходимо нормализовать лексику, т. е. заменить в тексте все синонимы введенного пользователем в запросе слова на это слово. Но в связи с возможностью омонимии возникла необходимость в разработке алгоритма ее устранения.

Этот модуль также может использоваться, например, в системах сравнения онтологий и в системах автоматизированной семантической разметки.

Данный алгоритм также может применяться для анализа именных групп, а не целых предложений. Например, в словосочетании «доход от продажи автомашины» морфологические характеристики каждого слова являются неоднозначными. Слово «доход» соответствует именительному или родительному падежу единственного числа, «продажи» – родительному единственного числа, именительному или родительному падежу множественного числа, слово «автомашины» соответствует родительному падежу единственного числа, именительному и родительному падежам множественного числа.

Однако, в документах официально-делового стиля в начале именной группы ставится главное слово, за предлогом «от» должно следовать слово в родительном падеже, слово без предлога относится к родительному-определительному. Таким образом, основными правилами анализа именных групп являются:

– главное слово в словосочетании – первое;

– если перед существительным имеется предлог, то морфологические характеристики слова определяются как пересечение возможных падежей данной словоформы и падежей, с которыми употребляется данный предлог;

– если существительное употреблено без предлога, то из всех возможных форм выбирается родительный падеж.

Такая модификация используется в программе семантического поиска информации в текстах правовых документов «Правотекст». Например, пусть пользователю необходимо найти только информацию об учете продаж продуктов питания, но не налоги доходов от этих продаж. Предполагается, что в онтологии пользователя есть понятия «продажа» и «продукты». Также в онтологии пользователя есть информация о том, что слово «продажа» является синонимом слова «реализация», и эти понятия являются подклассом слова "поставка". Слово «продукты» является синонимом слова «продукция», Пользователь вводит поисковый запрос, например, «Учет продаж продуктов». Запрос автоматически переводится на украинский язык с помощью переводчика Яндекс (если введен на русском языке), получаем "Облік продажів продуктів".

С учетом понятий онтологии исходный запрос заменяется на следующие: "Облік реалізацій продуктів", "Облік постачання продуктів", "Облік продажів продуктів", "Облік реалізацій продукції", "Облік постачання продукції", "Облік продажів продукції". После выполнения поиска всех вариантов будет получен результат "облік операцій з постачання власно виробленої продукції: молока, молочної сировини, молочних продуктів, м'яса, м'ясопродуктів, іншої продукції переробки тварин (шкур, субпродуктів, м'ясокісткового борошна), виготовленої з поставлених молока або м’яса в живій вазі сільськогосподарськими товаровиробниками (далі – продукція), і з постачання інших товарів/послуг, у тому числі продукції, виготовленої із сировини, визначеної у підпункті 1 цього пункту, ...".

Выводы

Точность данного алгоритма при обработке текста Налогового кодекса Украины близка к 100 %. При обработке менее формализованных текстов будут появляться неправильно распознанные и нераспознанные словоформы в причастных оборотах, а также прилагательных-существительных. Поскольку порядок слов в процессе снятия омонимии не учитывается, возможен правильный результат обработки непроективных конструкций.

Разработанный метод анализа ЕЯ позволяет также усовершенствовать онтологии ПрО, пополняя их знаниями, извлеченными из ЕЯ текста, с учетом снятия присутствующей в таких знаниях неоднозначности.

1. *Шкурко Е.В.* Синтаксическая омонимия и способы предупреждения ее возникновения. Ученые записки Таврического национального университета им. В.И. Вернад-

ского. Серия "Филология. Социальные коммуникации". 2011. Т.24 (63). № 2. Часть 2. С. 109–113.
2. *Гладкий А.В.* Синтаксические структуры. М.: Наука, 1985.
3. *Сокирко А., Толдова С.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка. http://aot.ru/docs/RusCorporaHMM.htm.
4. *Зеленков Ю.Г., Сегалович И.В., Титов В.А.* Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов. *Компьютерная лингвистика и интеллектуальные технологии.* Труды международного семинара Диалог, 2005. С. 188–197.
5. *Brill E.* Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics*. Vol. 21, N 4. P. 543–565.
6. *Лакомкин Е.Д., Пузыревский И.В, Рыжова Д.А.* Анализ статистических алгоритмов снятия морфологической омонимии в русском языке. http://aistconf.org/stuff/aist2013/submissions/aist2013_submission_33.pdf
7. *Анісімов А.В., Марченко О.О. Нагорний В.А.* Створення керуючого простору синтаксичних структур природної мови. Вісник Київського університету, серія: Фізико-математичні науки. Вип. 1, Київ. 2002.
8. *Марченко А.А., Никоненко А.А.* Контекстный семантический анализ текста. Система текстового мониторинга и качественного оценивания фокусного объекта. *Искусственный интеллект.* 2008. № 3. С. 809–813.
9. *Guarino N.* Formal Ontology in Information Systems. Formal Ontology in Information Systems. Proceedings of FOIS'98, 3-15, 1998.
10. *Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д.* Онтологии и тезаурусы: модели, инструменты, приложения. http://catscpp.googlecode. com/svn-history/r146/trunk/diploma/materials/ontologies_tesauruses.pdf
11. *Сучасна* українська літературна мова. За редакціїю М.Я. Плющ, 3-те видання, стереотипне. Київ. "Вища школа". 2001.
12. *Великий* тлумачний словник сучасної української мови. К., Перун. 2009.
13. *Лесько О.Н., Рогушина Ю.В.* Автоматизация семантической разметки естественно-языковых текстов. Материалы IX Международной научной конференции имени Т.А. Таран «Интеллектуальный анализ информации ИАИ-2009». Сб. тр. С. 247–253.

References

1. Shkurko E.V. (2011) Syntactic homonymy and ways to prevent its occurrence. In Scientific notes of Taurida national University. im V. I. Vernadsky. Series "Philology. Social communication", Vol. 24 (63), N 2. Part 2, P. 109–113. (In Russian).
2. Gladky A.V. (1985) Syntactic structure. M., Nauka. (In Russian).
3. Sokirko A. & Toldova S. (2005). Comparison of effectiveness of two methods of removing lexical and morphological ambiguity for the Russian language. http://aot.ru/docs/Rus Corpora HMM.htm. (In Russian).
4. Zelenkov Yu.G., Segalovich I.V., &Titov V.A. (2005) Probabilistic model of morphological disambiguity based on normalizing substitutions and positions of neighboring words. In Computer linguistics and intellectual technologies. Proc.of the international workshop Dialogue. P.188–197. (In Russian).
5. Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. In Computational linguistics, 21(4). P. 543–565.
6. Lakomkin E.D., Puzyrevskiy I.V. and Ryzhov D.A. (2013) Analysis of statistical algorithms of morphological homonymy in the Russian language. (In Russian). http://aistconf.org/stuff/aist2013/submissions/aist2013_submission_33.pdf
7. Anisimov A.V., Marchenko O.A. and Nagorny V.A. (2002) Creation of control space of syntactic structures of natural language. In Bulletin of Kiev University, series: Physical-mathematical science. Issue 1, Kiev. (In Ukrainian).
8. Marchenko O.O. and Nikonenko A.O. (2008) The Contextual Semantic Analysis of Natural Language Text. System of Text Monitoring and Qualitative Estimation of the Focus Object. In Artificial intelligence, N 3, P. 808–813. (In Russian).
9. Guarino N. (1998) Formal Ontology in Information Systems. In Formal Ontology in

Information Systems. Proceedings of FOIS'98. P. 3–15.
10. Dobrov B.V., Ivanov V.V., Lukashevich N. and Solovyev V.D. (2006) Ontologies and thesauri: models, tools, applications. (In Russian) http://catscpp.googlecode.com/svn-history/r146/trunk/diploma/materials/ ontologies_tesauruses.pdf
11. Modern Ukrainian literary language. Edited by M.J. Plusch (2001), 3rd edition, stereotyped, Kiev, High school. (In Ukrainian).
12. Big explanatory dictionary of modern Ukrainian language (2009). K. Perun. (In Ukrainian).
13. Lesko O.N. and Rogushina J.V. (2009) Automation of semantic markup of natural language texts. In Proc. of the IX international scientific conference named after T.A. Taran, "Intellectual analysis of information IAI-2009". P. 247–253. (In Russian).

Получено 18.04.2017

Об авторах:

Лесько Ольга Николаевна,
научный сотрудник.
Количество научных публикаций в украинских изданиях – 10.
orcid.org/0000-0002-5584-3799,

Рогушина Юлия Витальевна,
кандидат физико-математических наук,
старший научный сотрудник.
Количество научных публикаций в украинских изданиях – 140.
Количество научных публикаций в зарубежных изданиях – 30.
Индекс Хирша – 10.
http://orcid.org/0000-0001-7958-2557.

Место работы авторов:

Государственное учебно-научное учреждение "Академия финансового управления",
01014, г. Киев,
бульв. Дружбы Народов, 38.
E-mail: 12345o@i.ua,

Институт программных систем
НАН Украины,
03181, Київ-187,
проспект Академика Глушкова, 40.
Тел.: 066 550 1999.
E-mail: ladamandraka2010@gmail.com