ФОНЕТИКА
И ПРИКЛАДНАЯ ЛИНГВИСТИКА

УДК 81'366

КОРПУСНЫЙ МЕТОД АВТОМАТИЧЕСКОГО МОРФОЛОГИЧЕСКОГО АНАЛИЗА ФЛЕКТИВНЫХ ЯЗЫКОВ

О.И. Бабина, Н.Ю. Дюмин

A CORPUS METHOD FOR AUTOMATIC MORPHOLOGICAL ANALYSIS OF INFLECTIONAL LANGUAGES

O.I. Babina, N.Yu. Dyumin

Предложен метод автоматического морфологического анализа для языков флективного строя. Особенностью метода является работоспособность при отсутствии лексикона основ/псевдооснов, что достигается использованием корпуса текста на анализируемом языке.

Ключевые слова: автоматический морфологический анализ, автоматическая обработка текста, флективный язык, корпусные методы.

The article presents a method for automatic morphological analysis of inflectional languages. The analysis is based on the text corpus and it does not use any lexica which makes the method effective while being robust.

Keywords: automatic morphological analysis, natural language processing, inflectional language, corpus methods.

Введение

Морфологический анализ является базовым компонентом большого числа систем автоматической обработки текста (АОТ) на естественном языке, в том числе систем автоматического перевода, информационного поиска, автоматического извлечения информации. От эффективности работы морфологического анализатора во многом зависит эффективность работы всех последующих этапов и всей системы в целом.

Особенно важным является морфологический анализ флективных языков, в которых наибольшее затруднение вызывает фузия: флексия может иметь несколько грамматических значений, в отличие от словоизменительного аффикса в агглютинативных языках, который стремится к грамматической однозначности. Дополнительные трудности связаны с омонимичностью флексий в пределах одной словоизменительной парадигмы.

Проблемы морфологического анализа

Трудности автоматического морфологического анализа могут быть связаны 1) с особенностями конкретного языка и 2) с особенностями алгоритма автоматического анализа и его реализации.

Омонимия флексий. Омонимичными могут быть флексии: 1) принадлежащие одной словоизменительной парадигме, например, «лини-и» – флексия «и» (у существительных женского рода II мягкого склонения) может обозначать N_{SgGen}, N_{SgDat}, N_{SgPrep}, N_{PlNom}, N_{PlAcc}, 2) характеризующие одну лексико-грамматическую категорию, но при-

Бабина Ольга Ивановна, кандидат филологических наук, доцент, доцент кафедры лингвистики и межкультурной коммуникации, Южно-Уральский государственный университет (г. Челябинск). E-mail: olga_babina@mail.ru

Дюмин Никита Юрьевич, аспирант кафедры лингвистики и межкультурной коммуникации, Южно-Уральский государственный университет (г. Челябинск); научный руководитель – С.О. Шеремтьева, доктор филологических наук, доцент. E-mail: nikita.dyumin@gmail.com

Olga I. Babina, Candidate of Philology, Docent, Assistant Professor, Department of Linguistics and Cross-Cultural Communication, South Ural State University (Chelyabinsk). E-mail: olga_babina@mail.ru

Nikita Yu. Dyumin, post-graduate student, Department of Linguistics and Cross-Cultural Communication, South Ural State University (Chelyabinsk); supervisor – S.O. Sheremetyeva, Doctor of Philology, Docent. E-mail: nikita.dyumin@gmail.com

надлежащие разным словоизменительным парадигмам, например, флексия «и» может присутствовать также и в парадигмах первого («гени-и») и третьего («радост-и») склонений, 3) в некоторых случаях омонимичные флексии встречаются в парадигмах разных частей речи, например, «стекл-о» может иметь грамматическое значение $V_{PastSgNeu}$, либо N_{SgNeu}. Решение данной проблемы часто лежит за рамками собственно морфологического анализа: различные алгоритмы используют контекст, синтаксический анализ и т. п. для разрешения неоднозначности.

Внутренняя флексия. Данный тип флексий вызывает трудности при подходе с использованием словаря основ (необходимо указывать дополнительные словоформы) и особенно при подходах без использования словарей (необходимо описывать дополнительные сложные правила словоизменения).

Сложные слова. Слова, пишущиеся через дефис и требующие отдельного склонения для каждого компонента, например «монтажник-высотник», представляют трудность для систем всех видов и требуют отдельного рассмотрения. Кроме того, в языке могут существовать графические варианты слова, например «offline» – «off-line» – «off line».

Аналитические словоформы. Встречающиеся во многих языках аналитические словоформы могут представлять значительные трудности для разбора, так как компоненты словоформы могут быть разделены и находиться в разных позициях в предложении.

Большой объем лексического фонда языка. Несмотря на значительно возросшие вычислительные мощности современных компьютеров, работа со словарями основ и тем более словарями словоформ может оказаться проблематичной с технической точки зрения.

Подвижность лексического уровня языка. Появление новых слов, в частности терминов, уменьшает покрываемость лексического состава языка систем, основанных на словарях. Бессловарные системы в меньшей степени уязвимы в этом отношении.

Кроме этих наиболее общих и сложных проблем морфологического анализа имеются и другие, возникающие при реализации конкретных алгоритмов автоматического морфологического анализа.

Модели автоматического морфологического анализа

Варьируются и способы решения задач морфологического анализа. Выделяют два наиболее общих подхода: рационалистический и эмпирический. Первый подход использует лингвистические знания для анализа и синтеза словоформ, второй основан на эмпирических знаниях, например, на статистической модели текста. Однако в современной лингвистике, как правило, наиболее продуктивным оказывается гибридный подход, использующий преимущества обоих названных подходов.

По используемой базе знаний, лежащей в основе автоматического морфологического анализа, морфологические анализаторы можно подразделить на следующие типы:

- «словарные» системы:
  - системы с базой основ/лексем[1],
  - системы с базой словоформ[2];
- «бессловарные» системы (системы, не использующие лингвистических знаний).

Словарные системы с базой основ содержат словарь, хранящий список/списки неизменяемых частей слов (и иногда словосочетаний) подъязыка, к которому применяется морфологический анализ. Аддитивные правила позволяют «склеивать» элементы различных списков, формируя слово (при генерации), или раскладывать входное слово на части, идентифицируя их с элементами списков псевдоморфем (при анализе), причем порядок следования морфем задан. Такая элементно-комбинаторная модель чаще всего используется для флективных и агглютинативных языков; и порядок следования частей слова определяется как конкатенация классов морфем, например,

Префикс + *Корень* + *Суффикс* + *Окончание*,

где *Префикс* – список псевдоморфем, которые могут располагаться в начале слова, *Корень* – псевдоморфемы, которые следуют за префиксами т. д. Для описания знания о морфотактике, определяемой в системах автоматической обработки текста как порядок следования между элементами различных частей декларативной базы знаний (списков псевдоморфем), часто используется аппарат конечных автоматов[3]. В языках с трансфиксацией (например, в семитских языках) отдельно хранится база корней-радикалов и база шаблонов трансфиксов, показывающих каким образом согласные радикала распределяются между гласными трансфикса[4] (так называемые системы с «корневой» морфологией).

Наряду с декларативным знанием о составе морфем при применении элементно-операционного подхода база знаний также может содержать процедурное знание (например, об альтернациях, обусловленных фонологическими, лексическими или грамматическими причинами), то есть в базе хранится не только информация о морфемах, но и об операциях над ними. Операции, подвергающие морфему изменениям, представляются чаще всего продукционными правилами[5].

Системы, работающие на базе словаря основ, используют эмпирические методы с применением вероятностно-статистического подхода[6] для построения лексиконов суффиксов/псевдосуффиксов и основ/псевдооснов.

Системы с базой лексем включают в свой состав словарь лексических единиц в «начальной» форме. Эти единицы служат «эталоном» при проведении автоматического морфологического ана-

лиза словоформ по композиционным и/или продукционным правилам. Анализ словоформы признается верным в случае наличия сгенерированной по правилам системы леммы в словарной базе знаний[7]. В силу доступности достаточно представительных словарей в машиночитаемой форме такая модификация не является чрезмерно затратной с точки зрения накопления базы знаний.

Базы основ/лексем используются в морфологическом анализе для нормализации анализируемых словоформ. При наличии основ нормализация осуществляется в форме стемминга, а для модели с базой лексем в качестве нормализованной формы удобно принимать лемму.

В системах с базой словоформ понятия морфема как такового не существует. В таких системах реализуется словесно-парадигматический подход, в рамках которого грамматические значения описываются, оперируя непосредственно набором словоформ. Морфологический компонент сводится к перечислению словоформ, каждой из которых поставлено в соответствие грамматическая метка, описывающая лексико-грамматическое значение данной формы[8]. Преимуществом такого подхода является отсутствие ошибок при оформлении словоформы (так как словоформы не порождаются ни по каким аналитическим правилам, а хранятся в готовом виде). Это особенно актуально для слов с нерегулярной парадигмой. Неоднозначность, которая сохраняется, обусловлена лишь омонимией (омографией) полной словоформы (например,

*смет**а**на*_$N_{NomSgFem}$ – *см**е**тана*_$V_{SgFemShortPart}$,

*б**е**лки*_N_{GenSg} – *белк**и***_N_{NomPl}),

однако эта проблема неразрешима на уровне морфологии и требует учета синтаксического контекста.

Преимуществом словарных систем является лингвистическая обоснованность и прозрачность выбора варианта морфологического анализа.

Общим недостатком словарных систем является большой объем словаря, что, хотя и допустимо при современных мощностях компьютерной техники, однако делает систему громоздкой. Интуитивно кажется очевидным, что системы с базой основ/лексем с этой точки зрения занимают более выгодную позицию по сравнению с системами с базой словоформ. Впрочем, существует мнение, что реализация морфологического компонента на базе словаря словоформ сопоставима по степени избыточности информации с моделью морфологии на основе словаря основ, так как исчисление всех словоформ при лексикалистском подходе компенсируется избыточной базой знаний, описывающей отношения между морфемами в аддитивных и операционных моделях[9].

Другим недостатком является невозможность полного исчисления всех словоформ/лексем/основ в силу динамичности языка, что неизбежно приводит к сбоям работы такой системы при попытке ее применения к новому (для системы) текстовому материалу.

Наиболее существенным минусом словарной системы является трудоемкость составления лингвистической базы знаний, в значительной степени требующая привлечения ручного труда.

Бессловарные системы используют модели морфологии, основанные на применении математических методов машинного обучения. В таких системах используются различные алгоритмы (машины опорных векторов, EM-алгоритм, генетические алгоритмы, сети Кохонена и другие), позволяющие категоризировать словоформы на основе их автоматического вероятностно-статистического графематического анализа[10], результаты которого кладутся в основу формирования списка основ и суффиксов в данном языке.

Также способом применения в вычислительной морфологии методов машинного обучения является выявление категориальной соотнесенности словоформ. В этой задаче на основе выборки словоформ, корректно соотнесенных с лексико-грамматическим классом, алгоритм старается «научить» систему автоматизировать парадигматическую идентификацию словоформ[11]. Основой для машинного обучения могут служить как структурные, так и функциональные признаки словоформ.

Учитывая преимущества и недостатки различных методов, в нашем исследовании мы строим морфологический анализ, основываясь на гибридном статистико-рационалистическом подходе.

Корпусный метод и его компьютерная реализация

Предлагаемый нами корпусный метод автоматического морфологического анализа подразумевает наличие алгоритма, реализующего разбор, и базы лингвистических знаний. Алгоритм является универсальным для подмножества синтетических языков. В то же время база лингвистических знаний уникальна для каждого конкретного языка и содержит флексии (псевдофлексии) изменяемых частей речи данного конкретного языка. Таким образом, системы, реализующие предлагаемый метод, можно отнести к числу переходных словарно-бессловарных систем, что позволяет избежать проблем, связанных с подвижностью лексического состава языка и его объемностью.

Трудоемкость метода составляет сбор лингвистической базы знаний вручную, однако, в отличие от традиционных словарных систем, в нашем методе исчисляются лишь «конечные» классы морфологических элементов – флексии (точнее псевдоаффиксы). Флективные классы (кластеры словоизменительных псевдоаффиксов, репрезентирующих парадигмы слов) могут генерироваться автоматически[12], однако с целью сохранения точности категоризации словоформ этот компонент выполняется нами вручную. Как и в подходе морфологии цельнооформленных слов[13], где для автоматического выявления морфологических отношений базой является модель ментального лекси-

кона, в основе нашей модели лежат когнитивные принципы дифференциации и генерализации. При формировании базы флексий в нашей модели на основе статистического подхода выявляются сходные формальные показатели, выражающие одни и те же грамматические значения, и именно они составляют парадигму включенных в морфологически маркированную базу знаний манифестаций флективных классов.

Флексии в базе распределены в группы (парадигмы) по частям речи и, дополнительно, по грамматическим парадигмам склонения (или спряжения). Например, для русского языка имеются отдельные группы флексий для всех типов каждого из трех склонений существительного. Следует уточнить, что дистрибуция флексий по парадигмам в базе, в целом, основывается на данных практической грамматики анализируемого языка. Однако для повышения эффективности работы системы и для разрешения некоторых случаев омонимии окончаний применяются псевдофлексии, содержащие контактно расположенные по отношению к флексии символы основы. Таким образом, для базы флексий привлекаются данные анализа словников корпусов текстов.

Каждой флексии из базы лингвистических знаний приписана метка, включающая информацию о части речи, и грамматические характеристики, такие как род, число, падеж (для существительных), лицо, время, залог (для глаголов) и другие. Количественный и качественный состав грамматических категорий, входящих в метку, может варьироваться в зависимости от цели исследования и языка, для которого реализуется процедура автоматического морфологического анализа.

Помимо декларативных знаний, представленных флексиями (например, окончания правильного испанского глагола второго спряжения в форме Presente de Indicativo: «o», «es», «e» и т. д.), в базе знаний системы содержатся также процедурные знания, имеются сведения о возможных морфологических альтернациях[14]. Примером могут служить правила продукции парадигматических словоформ отклоняющихся испанских глаголов первой группы в форме Presente de Indicativo: «o:e→ie;1», «es:e→ie;1», «emos:e→e;1». Запись «o:e→ie;1» указывает на то, что для образования формы $V_{PresInd1pSg}$ необходимо добавить к основе аффикс «o» и, если последняя группа гласных основы «e», то заменить на «ie»; «1» указывает тип и место преобразования основы. На данный момент морфологический анализатор имеет три типа преобразований основы: 1) 0-преобразования, 2) преобразование последней группы гласных основы, 3) преобразование символов, расположенных контактно с флексией. Данные типы соответствуют структурам словоформ неизменяемых частей речи (1.0), изменяемых частей речи с флексией (1.1), изменяемых частей речи с флексией и альтернацией графем конечной части основы (1.2), изменяемых частей речи с внутренней флексией (1.3) и структурам изменяемых частей речи с флексией и альтернацией графем внутри основы (1.4).

R_0 (1.1)

R_0 I (1.2)

R_0 S I (1.3)

R_0 I R_1 (1.4)

R_0 S R_1 I (1.5)

В общем виде охватываемые структуры словоформ можно представить в следующем виде:

$$R_0 (S R_1^{?})^{?} I^{?}, \quad (2)$$

где R_0 – корень/инициальная часть корня, S – изменяемая часть основы, R_1 – финальная часть корня, I – аффикс, ? обозначает, что компонент может повторяться в формуле 0 или 1 раз.

Ограничение на структуру словоформы на данный момент продиктовано кругом морфологических явлений в рассматриваемых языках (русский и испанский), в дальнейшем, с увеличением количества языков возможно увеличение числа правил (способов) преобразования для покрытия большего числа явлений таких как, например, использование трансфиксов и циркумфиксов.

База лингвистических знаний, построенная подобным образом, может содержать несколько омонимичных флексий, кроме того часть флексий может совпадать по внешней форме с конечными частями словоформ, не являющимися флексиями или являющимися их частью, например, «скор-о*» вместо «скоро» (ср. «молок-о») или «hablab-a*» вместо «habl-aba» (ср. habl-a). Подобные проблемы разрешимы посредством использования корпуса текстов при автоматическом морфологическом анализе.

Для работы системы, помимо базы лингвистических знаний, необходим корпус текстов на анализируемом языке, причем, чем больше количество уникальных словоформ содержится в таком корпусе, тем выше точность морфологического анализа. Схема работы системы приведена на рисунке.

Использование лингвистических знаний и взаимодействие с корпусом происходит посредством алгоритма, который состоит из следующих шагов.

1. Для исследуемой словоформы из базы знаний подбираются подходящие флексии; флексии с

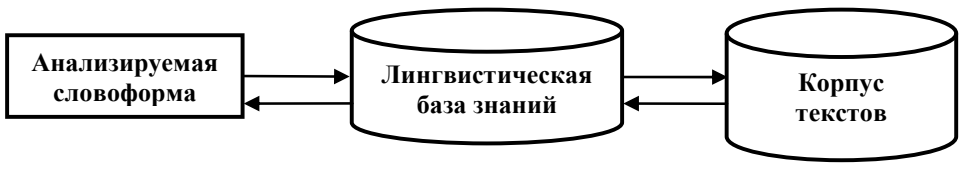

Схема автоматического морфологического анализа

указанием парадигм, к которым они относятся, формируют список парадигм-кандидатов.

2. Для каждой парадигмы-кандидата происходит воссоздание всех словоформ парадигмы с учетом возможных альтернаций, в качестве предполагаемой основы используется разность

W – I = S,

где W – анализируемая словоформа, I – предполагаемая флексия, S – предполагаемая основа (например, «слово» – «-о» = «слов-»).

3. Для каждой из сформированных парадигм определяется количество элементов этой парадигмы, зафиксированных в корпусе. Если это количество вхождений меньше установленного коэффициента для рассматриваемой парадигмы, то текущий кандидат удаляется.

4. Среди парадигм-кандидатов, относящихся к одной части речи, вычисляется парадигма с максимальным числом представленных элементов – такой кандидат считается верным с наибольшей вероятностью. Если несколько парадигм имеют одинаковое число вхождений в корпус, то учитываются дополнительный коэффициент парадигмы «слой», который фактически устанавливает приоритеты между парадигмами одной части речи (отбирается кандидат с наименьшим значением «слоя»).

5. После определения парадигмы словоформе приписывается соответствующая грамматическая помета, в случае омонимии в рамках одной парадигмы и омонимии частей речи выводятся все омонимы с соответствующими пометами (разрешение данного вида омонимии предлагается вынести за рамки собственно морфологического анализа).

Для проверки эффективности описанного метода разработана программа Co-MAIL (Corpus-based Morphological Analysis of Inflectional Languages).

В программной реализации морфологического анализа разработана база знаний для испанского и русского языков. При подключении соответствующей базы знаний входящая словоформа анализируется в соответствии с описанным алгоритмом и ей приписывается лексико-грамматическая информация (тэг). В процессе работы алгоритма для анализируемой словоформы строятся парадигмы-кандидаты. Среди возможных «хвостов» парадигмы отбираются те, которые хранятся в базе флексий подключенного языка. Например, для испанского глагола «entiendes» в базе испанских флексий хранится информация о флексии «-es» (2 парадигмы глагола – с альтернацией и без нее). Тогда для словоформы «entiendes» строится три гипотезы: в качестве парадигм-кандидатов могут быть восстановлены элементы двух парадигм, в которых встречается форма на «-es», а также принимается рабочая гипотеза о том, что данная словоформа принадлежит неизменяемой части речи. Таким образом, парадигмы-кандидаты строятся для основ «entiend-» (без учета альтернаций), «entend-» (с учетом альтернаций) и «entiendes» (предположение о неизменяемости слова). В табл. 1 можно видеть, какие элементы восстановленных парадигм уже имеются в корпусе (выделено жирным шрифтом), а какие отсутствуют (выделено курсивом). Следует отметить, что сгенерированные парадигмы могут содержать неверные словоформы, которые никогда не встретятся в корпусе (обозначены звездочкой в табл. 1). Так как парадигма, построенная с учетом альтернаций, имеет большее число представителей в корпусе, именно с этой парадигмой соотносится форма «entiendes» и данной словоформе приписывается метка $V_{PresIndAct2pSg}$.

После того как лексико-грамматическая идентификация словоформы произведена, возможно ее внесение в базу знаний системы, в случае, если эта словоформа не встретилась в корпусе. Таким образом, реализуется неконтролируемое самообучение системы – при следующем анализе отобранная парадигма будет представлена бо́льшим числом словоформ в корпусе.

Оценка алгоритма

Для оценки работы алгоритма нами был проведен эксперимент по выявлению эффективности морфологического анализа испанского глагола на ограниченных данных. Для эксперимента нами был собран корпус патентных текстов на испанском языке объемом около 55 тыс. словоупотреблений. Анализ текстов патентов выявил, что морфологическая парадигма глагола ограничивается в подавляющем большинстве случаев следующими формами:

Инфинитив	«hablar»
Наст. вр., 3 л., ед. ч.	«habla»
Наст. вр., 3 л., мн. ч.	«hablan»
Прост. прош. вр., 3 л., ед. ч.	«habló»
Сложн.прош.вр.	«ha/han hablado»
Буд.вр., 3л., ед.ч.	«hablará»
Буд.вр., 3л., мн.ч.	«hablarán»
Наст.вр., пасс.залог	«es/son hablado»
Герундий	«hablando»

Таблица 1

Сгенерированные элементы парадигм для словоформы «*entiendes*»

Предполагаемая основа	entiend-	entend-	entiendes
VPresIndAct1pSg	**entiendo**	**entiendo**	
VPresIndAct2pSg	*entiendes*	*entiendes*	
VPresIndAct3pSg	**entiende**	**entiende**	
VPresIndAct1pPl	**entiendemos*	**entendemos**	
VPresIndAct2pPl	**entiendéis*	*entendéis*	
VPresIndAct3pPl	**entienden**	**entienden**	
Неизменяемая ЧР			*entiendes*

Таблица 2

Список глагольных парадигм	Список неразобранных слов
accion: acciona accionado accionar	*abajo*
acept: acepta aceptan aceptar	*abertura*
acomod: acomoda acomodado acomodarse	*aberturas*
acopl: acopla acoplado acoplar acoplarse	*abierto*

Как можно видеть, часть этих форм аналитические, и формируются посредством вспомогательного глагола и причастия («hablado»). Рассматривая лишь синтетические формы, мы выявили флективные классы глагола, которые включают 7 синтетических форм. База знаний парадигм испанского глагола включает 12 флективных классов, в которые включены 3 спряжения невозвратного регулярно изменяющегося глагола, 3 спряжения возвратного регулярно изменяющегося глагола, а также парадигмы классов отклоняющихся глаголов (например, *sentir – siente, preferir – prefierte*).

В анализируемом корпусе текстов были выявлено 2354 различные словоформы. Алгоритм корпусного морфологического анализа был применен к списку выявленных словоформ. Результатом работы алгоритма явились список неразобранных словоформ и список глагольных основ с соответствующими им парадигматическими вариантами (начала списков приведены в табл. 2):

Список глагольных парадигм насчитывает 148 единиц, остальные 2206 не были разобраны. Далее нами вручную была проведена оценка эффективности алгоритма для решения задачи парадигматической идентификации глагольных словоформ. Для оценки нами использовались коэффициенты точности и полноты, заимствованные из теории информационного поиска. Подсчет производился по формулам:

$$P = \frac{v}{V_{tagged}};$$

$$R = \frac{v}{V_{corpus}},$$

где P – точность, R – полнота, v – количество корректно найденных парадигм глагола, V_{tagged} – общее количество найденных парадигм глагола, V_{corpus} – общее количество (различных) глагольных лексем в корпусе.

Согласно данным проведенного эксперимента, в целом было найдено 168 парадигм глагола, 5 из которых оказались неверными. Две парадигмы (с основами s- и d-) содержали наряду с правильными парадигматическими формами также служебные части речи de и se. Мы признали анализ форм, вошедших в парадигму, в целом правильным, так как эти единицы принадлежат служебным частям речи и проблема может решаться исчислением закрытых классов слов и их предварительным отфильтровыванием из списка словоформ до начала его анализа с помощью морфологического анализатора. 125 глагольных форм не были идентифицированы как глагольные. Таким образом, в процентном выражении точность работы алгоритма составила 97,02, а полнота – 58,33 %.

Анализ ошибок в работе алгоритма выявил в качестве основной причины, ухудшающей показатели точности алгоритма, синкретизм основ – от ряда основ могут формально порождаться аффиксальные формы, грамматические значения которых не составляют парадигму, соответствующую суффиксам, по которым эти формы сгенерированы. Например, для испанского языка при применении к основе s- парадигматических суффиксов глагола были сгенерированы формы «s-e», «s-er», каждая из которых имеет собственное лексическое значение и в целом они не формируют глагольную парадигму. Кроме того, отмечается совпадение флексий в парадигмах различных лексико-грамматических классов. Так, например, в качестве глагольной была признана парадигма, включающая ést-e (наст. вр., 3 л., ед. ч.), ést-a (сослаг. накл., 3 л., ед. ч.), хотя в действительности эти формы соответствуют указательным местоимениям.

Показатель полноты, в основном, невысок по двум причинам. Во-первых, наличие неправильных форм глагола, которые не были учтены при формировании списка парадигматических классов (например, «caer» – «caigo»). Другая, наиболее важная причина – это присутствие глагольной формы лишь в одном из своих парадигматических вариантов. Очевидно, применение алгоритма к б*о*льшему по объему корпусу увеличит вероятность встречаемости глагола в своих различных парадигматических вариантах, поэтому наращивание объема корпуса должно способствовать улучшению показателя полноты.

Заключение

Представленный метод использует положительные стороны двух подходов – словарного и бессловарного. С одной стороны, составленная лингвистическая база прозрачно и точно определяет границы парадигматических форм частей речи во флективных языках, с другой стороны, объем словарной базы знаний достаточно невелик и не требует больших трудозатрат для ее составления.

Оценка эффективности алгоритма, проведенная для одной части речи в ограниченной предметной области, показывает достаточно высокий результат, что дает основание полагать состоятельность и перспективность данного алгоритма. Дальнейшие шаги в развитии метода следует предпринять в направлении оценки работы алгоритмы в пространстве множества частей речи.

Другое направление – оценка возможности применения метода для других флективных и агглютинативных языков и разработка лингвистической базы знаний для них. Для возможности применения алгоритма к языкам с развитой внутренней флексией, очевидно, потребуется доработка лингвистической базы знаний.

[1] См., напр., Белоногов Г.Г., Кузнецов Б.А. Языковые средства автоматизированных информационных систем. М.: Наука, 1983. 288 с.; Шереметьева С.О. Методология минимизации усилий в инженерной лингвистике: дис. … д-ра филол. наук. СПб., 1997. 288 с.; Krovetz R. (2000). Viewing morphology as an inference process. In *Artificial Intelligence*, 118, 277–294; Dasgupta Sajib, Mumit Khan. (2004). Feature Unification for Morphological Parsing in Bangla. In *Proceedings of 7th International Conference on Computer and Information Technology (Dhaka, Bangladesh)*; Attia, Mohammed A. (2006). An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In *The Challenge of Arabic for NLP/MT Conference* (London, UK), 48–67; Paikens P. (2007). Lexicon-Based Morphological Analysis of Latvian Language. In *Proceedings of the 3rd Baltic Conference on Human Language Technologies* (Kaunas, Lithuania, 4–5 October, 2007), 235–240.

[2] См., напр., Sheremetyeva S., Nirenburg S., Nirenburg I. (1996). Generating Patent Claims From Interactive Input. In Proceedings of the 8th International Workshop on Natural Language Generation (Herstmonceux, Sussex, June 1996), 61–70; Mihalcea R. (2003). The Role of Non-Ambiguous Words in Natural Language Disambiguation. In *Proceedings of the Conference on Recent Advances in Natural Language Processing, RANLP 2003 (September 2003, Borovetz, Bulgaria)*. (http://www.cse.unt.edu/~rada/ papers/mihalcea.ranlp03.pdf) и др.

[3] См. Koskenniemi K. (1990). Finite-State Parsing and Disambiguation. In Proceedings of COLING-90, Vol. 2, 229–232; Lauri Karttunen. (1993). Finite state constraints. In John A. Goldsmith (ed.), *The Last Phonological Rule*, Chicago: University of Chicago Press, 173–194; Abney S. (1996). Part-of-speech tagging and partial parsing. In G.K. Church, S. Young (ed.), *Corpus-based methods in language and speech*. Kluwer academic publishers, Dordrecht; Dasgupta Sajib, Mumit Khan. (2004). Feature Unification for Morphological Parsing in Bangla. In *Proc. 7th ICCIT*; Beesley K.R., Karttunen L. (2003). Finite State Morphology. Stanford, CA: CSLI Publications, 2003. 505 p.; Attia, Mohammed A. (2006). An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In In *The Challenge of Arabic for NLP/MT Conference* (London, UK), 48–67; Köprü Selçuk, Jude Miller. (2009). A Unification Based Approach to Morphological Analysis and Generation of Arabic. In *CAASL3: Third Workshop on Computational Approaches to Arabic- Script-based Languages* (Ottawa, Canada, August 26, 2009).

[4] Beesley K.R. 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In *ACL Workshop on Arabic Language Processing: Status and Perspective* (Toulouse, France), 1–8; Habash Nizar, Owen Rambow, and George Kiraz. (2005). Morphological Analysis and Generation for Arabic Dialects. (http://www1.cs.columbia.edu/~rambow/papers/magead-ws05.pdf)

[5] Mohammed Attia (2006). An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In *The Challenge of Arabic for NLP/MT Conference* (London, UK), 48–67.

[6] Шереметьева С.О., Ниренбург С. Эмпирическое моделирование в вычислительной морфологии // НТИ. 1996. № 7; Белоногов Г.Г. Итоги науки и техники. Серия «Информатика». 1984. № 8.

[7] Напр., Krovetz R. (2000). Viewing morphology as an inference process. In *Artificial Intelligence*, 118, 277–294.

[8] Напр., Sheremetyeva S., Nirenburg S., Nirenburg I. (1996). Generating Patent Claims From Interactive Input. In Proceedings of the 8th International Workshop on Natural Language Generation (Herstmonceux, Sussex, June 1996), 61–70.

[9] См. Подробнее Kirby J. (2006). Minimal Redundancy in Word-Based Morphology. (http://home.uchicago.edu/~jkirby/docs/ morph_rewrite.pdf)

[10] Kazakov D. (1997). Unsupervised Learning of Naïve Morphology with Genetic Algorithms. In *Workshop Notes of the ECML/MLnet workshop on empirical learning of Natural Language Processing Task* (Prague, Czech Republic, April 1997), 105–112; Goldsmith J. (2001). Unsupervised Learning of the Morphology of a Natural Language. In *Computational Linguistics*, 27(2), 153–198.

[11] Например, Zhao, Jian, and Xiao-Long Wang. (2002). Chinese POS Tagging Based on Maximum Entropy Model. In *Proceedings of the First International Conference on Machine Learning and Cybernetics,* Beijing, 4–5 November 2002, 601–605; Masuyama Takeshi, and Hiroshi Nakagawa. (2004). Two Step POS Selection from SVM Based Text Categorization. In *IEICE Trans. Inf. & Syst*. Special Issue on Information Processing Technology for Web Utilization, Vol. E87-D, No. 2, February 2004. (http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/academic-res/masuyama-ieice-04.pdf);
Klami Mikaela, and Krista Lagus. (2006). Unsupervised Word Categorization Using Self-Organizing Maps and Automatically Extracted Morphs. In E. Corchado et al. (eds.) IDEAL 2006, LNCS 4224. – Berlin;Heidelberg: Springer-Verlag, 2006, 912–919.

[12] Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. In *Computational Linguistics*, 27(2), 153–198.

[13] Ford A., & Singh R. (1991). Propedeutique Morphologique. *Folia Linguistica,* 25 (3–4), 549–575; Neuvel Sylvain (2002). Whole Word Morphologizer: Expanding the Word-Based Lexicon: A Nonstochastic Computational Approach. In *Brain and Language* 81, 454–463.

[14] Бабина О.И., Дюмин Н.Ю. Нестрого аддитивный подход к автоматическому морфологическому анализу флективных языков // Материалы 5-й Междунар. науч.-практ. конф. «Наука и современность-2010». Секция «Филологические науки». – Новосибирск: Центр развития научного сотрудничества, 2010, 12–17.

Поступила в редакцию 10 декабря 2011 г.