

АВТОМАТИЧЕСКИЙ МОРФОЛОГИЧЕСКИЙ АНАЛИЗ ФЛЕКТИВНЫХ ЯЗЫКОВ

О.И. Бабина, Н.Ю. Дюмин

Для языков флективного типа характерным является широкое использование аффиксов в процессе словообразования и словоизменения, что делает их схожими с агглютинативными языками. В названных типах языков форма слова выражает 1) лексическое значение, которое приходится преимущественно на основу и 2) грамматическое значение, которое приходится на аффикс (флексию). Принципиальное отличие флективных языков от агглютинативных – полисемия / омонимия словоизменятельных аффиксов: с одной стороны, флексия может интегрировать в себе ряд грамматических значений, с другой стороны, несколько рядов грамматических значений могут быть выражены омонимичными флексиями. Кроме того, для флективных языков характерно явление фузии. В агглютинативных языках аффиксы тяготеют к грамматической однозначности.

Морфологический анализ позволяет выделить в структуре слова морфемы, в том числе и флексии. В компьютерной морфологии под морфологическим анализом понимается «процедура, в результате которой из формы, внешнего оформления слова в тексте можно получить сведения о самых различных уровнях языковой структуры» [1, с. 61]. То есть в ходе автоматического морфологического анализа можно идентифицировать словоформу и классифицировать ее по принадлежности определенному морфологическому классу.

Выделяется два вида идентификации словоформ – синтагматическая и парадигматическая [2, с. 94]. Синтагматическая идентификация применяется при автоматическом анализе дискурса, в ходе которого вычленяются словоформы из необработанного текста. В ходе парадигматической идентификации словоформы соотносятся с их лексическими инвариантами и, таким образом, группируются в функциональные классы. Каждая словоформа соотносится с соответствующей ей словарной единицей, и проводится содержательная интерпретация формального различия между текстовой словоформой и инвариантом словарной единицы.

В формальной морфологии в ходе парадигматической идентификации для выявления структуры словоформы с целью последующей интерпретации выделяют три класса моделей [3]: элементарно-комбинаторную (Item and Arrangement), элементарно-операционную (Item and Process) и словесно-парадигматическую (Word and Paradigm). Последняя модель оперирует словоформой как минимальной единицей и

потенциально универсальна для различных типов языков; эта модель применима для построения алгоритмов морфологического анализа языков различных морфологических типов, а также машинного обучения морфологии (см., например, [4]). Первые две модели концептуально более просты, и в большей степени применимы лишь для флективных языков (так как во флективных языках имеется достаточно четко выраженное соответствие между флексией и грамматическим значением соответствующей словоформы).

Во флективных языках, таком как русский, грамматическое значение, как правило, несет в себе аффиксальная морфема, то есть формальный показатель грамматического значения условно соответствует последовательности букв в конце слова. Наборы суффиксов, имеющих определенное грамматическое значение, регулярны для различных словоформ. Например,

<u>инфинитив</u>	<u>1л. ед.ч. наст.вр.</u>	<u>2л. ед.ч. наст.вр.</u>	и т.д.
‘кур-и-ть’	‘кур-ю’	‘кур-и-шь’	и т.д.
‘посел-и-ть’	‘посел-ю’	‘посел-и-шь’	и т.д.
‘помн-и-ть’	‘помн-ю’	‘помн-и-шь’	и т.д.
‘вер-и-ть’	‘вер-ю’	‘вер-и-шь’	и т.д.

В примере отчетливо прослеживается систематический характер изменения грамматического значения слова путем добавления к основе идентичных для различных словоформ финалий с соответствующим грамматическим значением. Такое поведение словоформ дает основания для использования в автоматическом морфологическом анализе аддитивного подхода.

Однако строго аддитивный подход, основывающийся на использовании элементарно-комбинаторной модели (которая заключается в механическом присоединении сегментов слова друг к другу), оказывается недостаточно эффективным, так как во флективных языках нередко случаи фонетические преобразования морфем, что может приводить к формированию таких парадигматических форм, как ‘крас-и-ть’ – ‘краш-у’, ‘воз-и-ть’ – ‘вож-у’, ‘груз-и-ть’ – ‘грузж-у’, ‘лат-и-ть’ – ‘лач-у’, ‘крут-и-ть’ – ‘круч-у’, ‘груст-и-ть’ – ‘грущ-у’, ‘прост-и-ть’ – ‘прощ-у’ и др.

Очевидно, что, в силу влияния таких явлений как чередующиеся согласные, беглые гласные и т.п., использование классического разбиения слова на морфемы не укладывается в рамки данного подхода. Традиционно, в системах автоматической обработки текста существует два решения этой проблемы. Во-первых, проведение морфологического анализа в рамках элементарно-операционной модели через описание процедурного знания о морфологических преобразованиях, которые необходимо провести над морфемой. Например, следующее правило в рамках элементарно-операционной модели $\text{Verb}_{\text{pres},1,\text{sing}} \rightarrow \text{Stem} | 'm' \#_1 'ч' + 'у'$

указывает на то, что для образования первого лица, единственного числа в настоящем времени таких глаголов как, например, *‘крутить’*, *‘платить’*, *‘катить’* и т.д., необходимо заменить в основе первую с конца букву *‘т’* на *‘ч’* и добавить окончание *‘у’*.

Другим решением представляется стратегия, в ходе которой в рамках элементарно-комбинаторной модели слово представляется не как последовательность соположенных морфем, а как конкатенация *квазиморфем*, под которыми понимается множество таких сегментных знаков, квазипредставимых по означаемому, таких что их означаемое тождественно, а сами эти знаки распределены по стандартным правилам, действующим внутри рассматриваемой словоформы [5, с. 59, с. 255]. Тогда весь набор лексем может быть разбит на *флексивные классы* [6, с. 119], или *типовые парадигмы* [2, с. 102], в каждую из которых входят лексем, образующие парадигматические формы по формуле *квазиоснова + квазисуффикс*, причем набор квазисуффиксов с соответствующим грамматическим значением идентичен для всех квазиоснов одного флексивного класса. Тогда, например, такие квазиосновы как *‘пла-’*, *‘кру-’*, *‘ка-’* и т.п. образуют словоформы морфологической парадигмы с помощью квазисуффиксов *‘-тить’*, *‘-чу’*, *‘-тишь’* и т.д.

Как и большинство исследователей, для решения задачи автоматического морфологического анализа русскоязычного текста мы придерживаемся элементарно-комбинаторной модели (оперирующей квазиморфемами в качестве составных единиц словоформы), так как эта модель интуитивно понятна и более проста в реализации алгоритмически. С использованием данной модели нами составлена формальная процедура парадигматической идентификации словоформ, в ходе которой проводится поиск в русскоязычном корпусе текстов словоформ, принадлежащих одной морфологической парадигме, и их группировка.

База знаний, необходимая для реализации данного алгоритма, включает в себя набор списков квазисуффиксов, где каждый список соответствует одному флексивному классу. Словоформы, имеющие финалии, описанные в списке квазисуффиксов базы знаний, рассматриваются как кандидаты для формирования морфологических парадигм лексем.

В процессе поиска парадигматических вариантов лексем выстраиваются гипотезы о принадлежности каждой словоформы-кандидата одному или нескольким флексивным классам. По причине дивергентной омонимии некоторых квазисуффиксов и включающих их в свой состав словоформ возможна неоднозначность – одна и та же словоформа может потенциально принадлежать нескольким парадигмам. Так, квазисуффикс *–и* является частью флексивных классов, представляющих существительное (напр., форма *‘корабли’*), прилагательное (напр., форма *‘редки’*) и глагол (напр., форма *‘выходи’*).

Поэтому для автоматического определения грамматического значения соответствующей словоформы проводится сравнительный анализ данной словоформы с другими лексическими единицами в тексте, и на базе вероятностных коэффициентов определяется принадлежность словоформы той или иной парадигме (флективному классу). В основе сравнения лежит гипотеза о том, что в корпусе количество различных словоформ одной лексемы, которые могут быть разложены на квазисуффиксы, соответствующих «правильному» для данной лексемы флективному классу, значительно больше, чем количество словоформ, группа которых могла бы быть отнесена к «неверному» для анализируемой словоформы флективному классу. Так, словоформа *‘корабл-и’* по вероятностному коэффициенту, скорее всего, идентифицируется как существительное (во множественном числе, именительном падеже), так как в корпусе количество форм с квазисуффиксами одного из флективных классов существительного (таких как *‘корабл-ей’*, *‘корабл-я’* и т.д.) будет больше, чем количество словоформ, которые обеспечили бы принятие решения об отнесении этой формы к классу прилагательных или глаголов (так, в корпусе, несомненно, не будет словоформ **‘корабл-ий’*, **‘корабл-еишь’* и т.п.).

Таким образом, в основе механизма анализа морфологического состава слова во флективном языке лежат корпусный и статистико-рационалистический подходы. При поиске парадигматических вариантов определенную роль играет объем корпуса, на материале которого выявляются парадигмы лексем. В силу диффузности информации, извлекаемой из корпуса текстов, точность вероятностных коэффициентов в значительной степени зависит от объема исследовательского корпуса.

Библиографический список

1. Марчук, Ю.Н. Компьютерная лингвистика: учебное пособие / Ю.Н. Марчук. – М.: АСТ: Восток – Запад, 2007. – 317 с.
2. Коваль, С.А. Лингвистические проблемы компьютерной морфологии / С.А. Коваль. – СПб.: Изд-во С.-Петербург. ун-та, 2005. – 151 с.
3. Matthews, P.H. Morphology, 2nd ed. / P.H. Matthews. – Cambridge: Cambridge University Press, 1991. – xii, 251 p. – (Cambridge textbooks in linguistics).
4. Neuvel, S. Unsupervised Learning of Morphology without Morphemes / Sylvain Neuvel, Sean A. Fulop // Morphological and Phonological Learning: Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON) (Philadelphia, July 2002). – Philadelphia: ACL, 2002. – Pp. 31-40.
5. Мельчук, И.А. Курс общей морфологии. Т. IV / И.А. Мельчук; пер. с фр. Е.Н. Саввиной под общ. ред. Н.В. Перцова. – М.; Вена: Языки славянской культуры: Венский славистический альманах, 2001. – 584 с.
6. Белоногов, Г.Г. Языковые средства автоматизированных информационных систем / Г.Г. Белоногов, Б.А. Кузнецов. – М.: Наука, 1983. – 288 с.