К.К. Боярский, Е.А. Каневский

РАЗРАБОТКА ИНСТРУМЕНТАРИЯ ДЛЯ ПОЛУАВТОМАТИЧЕСКОЙ МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ ТЕКСТОВ[1]

Как правило, разметка заключается в приписывании текстам и языковым единицам специальных меток. При *морфологической разметке* метки включают не только признак части речи, но и признаки грамматических категорий, свойственных данной части речи. Процедура морфологической разметки является в большой степени отдельной областью исследований и имеет результат, обладающий самостоятельной ценностью – текст, содержащий правильно и однозначно расставленные морфологические маркеры (тэги).

Однако автоматический анализ естественного языка, в том числе и морфологический, небезошибочен и многозначен – часто он дает несколько вариантов анализа для одной языковой единицы. В этом случае говорят о грамматической омонимии, что приводит к порождению множества вариантов анализа. В связи с этим обстоятельством задачу снятия морфологической неоднозначности можно представить следующим образом: для каждого слова в тексте необходимо выбрать из всего списка возможных тэгов единственно правильный.

Необходимость преодоления неоднозначности морфологической интерпретации словоформ при автоматическом анализе текста признается многими лингвистами, решающими прикладные задачи. Снятие неоднозначности результатов анализа является одной из важнейших и сложнейших задач компьютерной лингвистики.

Снизить уровень неоднозначности можно, используя либо обучаемые программы, либо семантико-синтаксические анализа-

[1] Работа выполнена при поддержке гранта РГНФ № 07-04-00464а.

торы[1]. В первом случае требуются предварительные громадные затраты ручного труда на составление размеченных корпусов. Во втором случае проблема упирается в отсутствие качественных анализаторов, работа которых на реальных текстах (а не тестовых примерах) оставляет желать лучшего.

Авторами был предложен способ снятия морфологической неоднозначности, основанный на сравнении результатов разбора текста двумя программами, использующими разные принципы[2]. В качестве первой программы взят морфологический анализатор системы *Диалинг*[3], второй – морфолого-семантический анализатор, построенный по технологии *SemLP*[4]. Однако, как показали эксперименты, эти программы по-разному определяют границы предложений в зависимости от трактовки знаков препинания, что значительно усложняет сопоставление результатов разбора.

В связи с этим была разработана программная оболочка для снятия морфологической неоднозначности *KillAmbig*. За основу берется выходной файл системы *Диалинг*, а единицей сравнения является предложение в границах, определяемых этой системой. Каждое предложение затем разбирается анализатором *SemLP*. Далее производится сравнение результатов работы обеих программ и по каждому варианту морфологического разбора подсчитывается некоторый условный вес совпадения. Для каждого сло-

---

[1] *Сокирко А. В, Толдова С. Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // Интернет-математика 2005. Автоматическая обработка веб-данных. М.: Яндекс, 2005. С. 80–94.

[2] *Боярский К.К., Захаров В.П., Каневский Е.А*. Снятие неоднозначности морфологической разметки корпусов русских текстов // Труды международной конференции «Компьютерная лингвистика – 2006». СПб.: СПбГУ, 2006. С. 70–74.

[3] URL: **www.aot.ru**

[4] URL: **www.semlp.com**

ва подсвечивается тот вариант, которому соответствует максимальный вес, и этот вариант принимается за правильный.

Рассмотрим работу *KillAmbig* на примере анализа предложения *Ковры были в пятнах* (И.А. Гончаров, «Обломов»). Результат разметки этого предложения системой *Диалинг* (см. рис. 1):

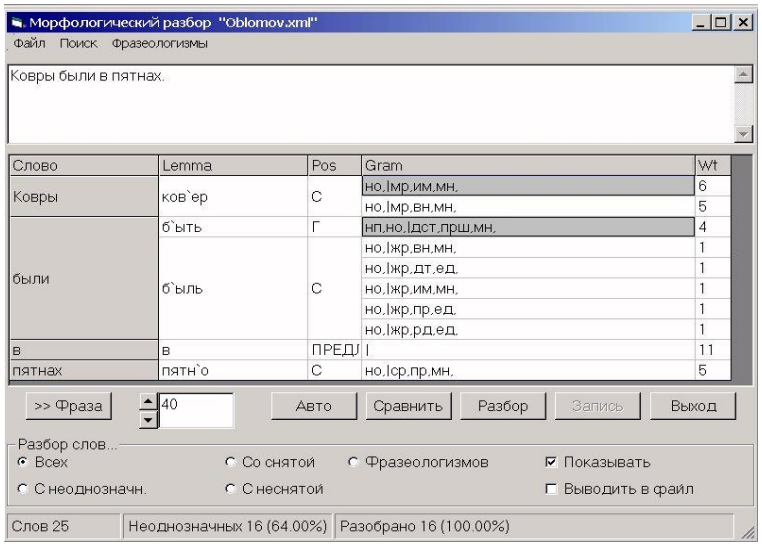

Рис. 1. Снятие морфологической неоднозначности

Ковры – *ковер*, существительное муж. рода, им./вин. падежа, мн. числа, всего два варианта;

были – мн. число прош. времени от глагола *быть*, или род./дат./пр. падежи ед. числа, или им./вин. падежи мн. числа от существительного *быль*, всего шесть вариантов;

в – предлог, разобран однозначно;

пятнах – существительное ср. рода, пр. падежа, мн. числа, разобрано однозначно.

Эта же фраза разбирается анализатором *SemLP*, который выдает им. падеж для слова *ковры* и глагол для слова *были*, таким образом, морфологическая неоднозначность оказывается полностью снятой.

Еще пример из того же произведения. Диалинг разбирает предложение *Старые господа умерли, фамильные портреты остались дома и, чай, валяются где-нибудь на чердаке* так:

Старые – им./вин. падеж мн. числа от прилагательного *старый* или существительного *старое*, четыре варианта;

господа – им. падеж мн. числа от слова *господин*, либо род./вин. падежи ед. числа от слова *господь*, три варианта;

фамильные – им./вин. падежи мн. числа, два варианта;

портреты – им./вин. падежи мн. числа, два варианта;

дома – род. падеж ед. числа либо им./вин. падежи мн. числа от слова *дом*, или наречие *дома*, четыре варианта;

и – междометие либо союз, два варианта;

чай – им./вин. падежи от слова *чай*, либо повел. наклонение от глагола *чаять*;

остальные слова разобраны однозначно.

SemLP, разбирая эту фразу, однозначно определяет основную форму и падежи для слов *старые*, *господа*, *фамильные*, *портреты*; показывает, что *дома* – это наречие, а, кроме того, правильно находит, что *чай* здесь является вводным словом, чего нет в разборе *Диалинга*.

Использование двух независимых анализаторов позволяет в автоматическом режиме резко снизить неоднозначность разбора. Если *Диалинг* дает уровень неоднозначности порядка 50%, то после работы *KillAmbig* он не превышает 8–10%. Оставшаяся неоднозначность связана, преимущественно, с двумя факторами. Во-первых, это трудность сравнения некоторых слов. Так *Диалинг* во фразе *…человек лет тридцати двух-трех от роду…* словосочетание *двух-трех* трактует как единое слово, а *SemLP* – как два слова. Однако алгоритм сравнения постоянно совершенствуется и неоднозначностей такого рода становится все меньше.

Второй фактор – ошибки того или другого анализатора. Например, *Диалинг* в предложении *Э-э-э! слишком проворно!* считает, что *Э-э-э* – это существительное. Имеются свои недостатки и у *SemLP*, в основном связанные с определением служебных слов, местоименных прилагательных и т.д.

Очевидно, что анализ вариантов разбора, не совпавших у двух анализаторов, является мощным инструментом отладки и совершенствования самих анализаторов.

Особым режимом работы программы *KillAmbig* является анализ словосочетаний и фразеологизмов. Дело в том, что *Диалинг* производит пословный разбор, а *SemLP* анализирует словосочетание в целом. Некоторые из них являются устойчивыми словосочетаниями или составными словами: *бить баклуши*, *попасть впросак*, *как будто* и т.п.[1] Другие же появились вследствие затруднений при выборе анализатором правильной языковой конструкции: *близкий к*, *в момент*, *вести себя*. Поэтому автоматическое сопоставление результатов в большинстве случаев просто невозможно.

Дойдя до неизвестного фразеологизма, система переходит из автоматического в интерактивный режим, в котором пользователь указывает правильные варианты разбора каждого слова (из вариантов, предложенных *Диалингом*), а также признак изменяемости слов. Слова, образующие фразеологизм, могут не изменяться вообще (*во всяком случае*). Может изменяться только первое слово, причем это измененяемое слово может быть глаголом (*покраснеть до ушей*), существительным (*жизнь на водах*), прилагательным (*единственный в мире*) или числительным (*один из них*). В случае согласования пары прилагательное – существительное могут изменяться оба слова (*эолова арфа*). Выделение того или иного словосочетания в качестве фразеологизма задается словарем. Их количество довольно велико, так, например, в романе «Обломов» 440 фразеологизмов встречаются более 3000 раз[2].

---

[1] *Кузнецов С.А.* Большой толковый словарь русского языка. СПб.: Норинт, 1998.

[2] *Захаров В.П., Каневский Е.А.* Роман И.А. Гончарова «Обломов» сквозь призму современной грамматики // Прикладная лингвистика в науке и образовании. Материалы научно-практической конференции 27–28 марта 2008. СПб.: Лемма, 2008. С. 74–80.

Разобранные фразеологизмы помещаются в специальную базу данных, в которой хранятся как уточненная характеристика самого фразеологизма, так и морфологические характеристики всех слов, входящих в его состав (рис. 2, разбор фразеологизма *дело в том*). Это позволяет при повторном появлении в тексте этого фразеологизма обрабатывать его уже автоматически. После этого уровень неснятой морфологической неоднозначности падает до 4–6%.

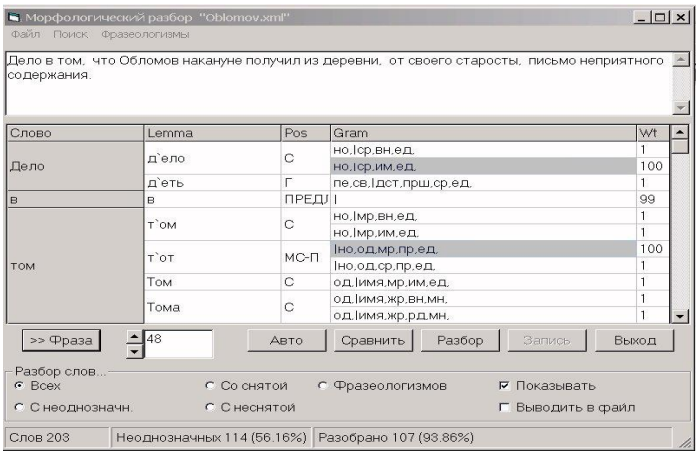

Рис. 2. Разбор фразеологизма

Таким образом, выбранный способ снятия неоднозначности морфологической разметки путем сравнения результатов работы двух анализаторов текста, построенных на разных принципах, позволяет в автоматическом режиме снизить уровень неоднозначности до нескольких процентов, а в интерактивном режиме служит мощным средством контроля и отладки.