Проблемы бионики
1999. Вып. 50. С. 179-184
ХТУРЭ
УДК 519.767

Д.Э.СИТНИКОВ, Н.Ф. ХАЙРОВА, Н.В. ШАРОНОВА

МОДЕЛИРОВАНИЕ СЕМАНТИКО-СИНТАКСИЧЕСКИХ ОТНОШЕНИЙ ГРАММАТИЧЕСКИХ СЛОВОСОЧЕТАНИЙ

При морфологическом анализе системы машинного перевода (МП) каждой словоформе предложения приписывают морфологическую информацию (МИ). Так как словоформы анализируются вне связи с контекстом, практически каждая из них обладает морфологической омонимией, в результате чего ей приписывается целый комплекс морфологической информации (КМИ), т.е. набор возможных альтернативных вариантов морфологической информации. В системах автоматической обработки текстов на естественном языке (ЕЯ) морфологическая омонимия обычно снимается на этапе синтаксического и даже семантического анализа. В современных промышленных системах МП, работающих с большими словарями в широких областях знаний, семантический анализ представлен только семантическими фильтрами. Алгоритм глубинного синтаксического анализа достаточно сложен и в силу этого далеко не всегда реализуется в полной мере [1].

Предлагается использовать метод компараторной идентификации, описанный в [3], на материале семантико-синтаксических отношений грамматических словосочетаний, применение которого, на первом этапе синтаксического анализа, позволит снять значительную часть морфологической омонимии.

В результате морфологического анализа каждой словоформе предложения приписывается морфологическая информация, которая является многозначной (или морфологически омонимичной). Однако при образовании семантико-синтаксической связи между рядом стоящими словоформами в предложении не все МИ могут быть связаны отношением. Следовательно, необходимо выявить и математически описать закономерность образования связей между двумя рядом стоящими словоформами в предложении [2].

Рассмотрим множество словоформ $M=\{m_1, \ldots, m_n\}$, где n — количество словоформ в словаре системы. Словоформы из множества M образуют словосочетания, т.е. между словоформами устанавливаются семантико-синтаксические связи, которые можно выразить формально, используя основные методы компараторной идентификации лингвистических объектов [4].

Грамматическое словосочетание можно представить в виде: $m_i * m_j$, где $m_i, m_j \in M$, а знак $*$ — обозначает, что между словоформами установлены определенные семантико-синтаксические связи. В проективных предложениях индоевропейских языков (а системы МП работают только с

проективными предложениями) если две словоформы связаны друг с другом, то связи между ними образуют три типа грамматического подчинения: согласование, управление и примыкание.

На множестве M введем систему предикатов S так, чтобы любой предикат $P(q_m) \in S$ обращался в *1* на множестве словоформ с какой-то определенной морфологической информацией и был равен *0* в противном случае. Таким образом, множество предикатов S можно сопоставить с комплексом морфологической информации, приписываемой словоформе словосочетания на этапе морфологического анализа. Каждой словоформе m_i из M соответствует некоторый предикат $P(q_m) \in S$, равный *1* при подстановке комплекса морфологической информации, приписанного на предыдущем этапе анализа конкретной словоформе m_i. Следовательно, каждому элементу m взаимно-однозначно соответствует определенный одноместный предикат, задающий комплекс морфологической информации словоформы словосочетания. Операция соединения двух словоформ из M, комплексы морфологической информации которых заданы предикатами $P(q_m) \in S$ и $P(q_n) \in S$, характеризуется определенной семантико-синтаксической связью, которая определяет тип грамматического подчинения в словосочетании. В результате семантико-синтаксического согласования двух рядом стоящих словоформ получаем множество связей между КМИ, другими словами — множество возможных семантико-синтаксических связей в различных типах грамматического подчинения в словосочетаниях. Таким образом, между КМИ рядом стоящих словоформ предложения существует бинарное отношение, которое является подмножеством декартового произведения этих комплексов.

Это бинарное отношение можно представить с помощью двухместного предиката $P(q_m, q_n)$, при этом для любых q_m, q_n

$$P(q_m, q_n) \rightarrow P(q_m) \bullet P(q_n), \qquad (1)$$

где $\bullet$ — операция конъюнкции предикатов.

Предположим, что возможность согласования комплексов морфологической информации не зависит от того, к каким словоформам они относятся. Тогда на декартовом произведении множества $S * S$ можно задать предикат $\gamma_i(q_m, q_n)$, принимающий значение *1*, если морфологические информации словоформ q_m и q_n связаны при данном типе грамматического подчинения, и значение 0 — в противном случае. Подмножество согласуемой морфологической информации практически никогда не совпадает с декартовым произведением всех возможных связей. Те МИ рядом стоящих словоформ, которые не согласуются при данном типе грамматического подчинения, исключаются из формулы (1) множителем $\gamma_i(q_m, q_n)$, $i = 1,2,3$ (согласование , управление, примыкание). Таким образом, бинарное отно-

шение на множестве рядом стоящих словоформ предложения для всех типов грамматического подчинения может быть задано формулой

$$P(q_m) * P(q_n) = \gamma_i(q_m, q_n) \bullet P(q_m) \bullet P(q_n), \qquad (2)$$

где знак * обозначает операцию соединения комплексов морфологической информации словоформ (операцию связи МИ двух рядом стоящих словоформ, т.е. знак * указывает на то, что две рядом стоящие словоформы связаны между собой семантико-синтаксической связью). Действительно, логическое произведение предикатов $P(q_m)$ и $P(q_n)$ описывает всевозможные связи между двумя рядом стоящими словоформами в предложении, а предикат $\gamma_i(q_m, q_n)$ исключает часть связей, которые не реализуются в данном типе грамматического подчинения анализируемого языка.

Рассмотрим работу данной модели на примере предложений русского языка. Выберем простейшую систему морфологических категорий и их значений, состоящую из части речи и наиболее существенных морфологических признаков. Если система МП настраивается на обработку сложных текстов, то систему грамматических категорий можно расширить без изменения алгоритма.

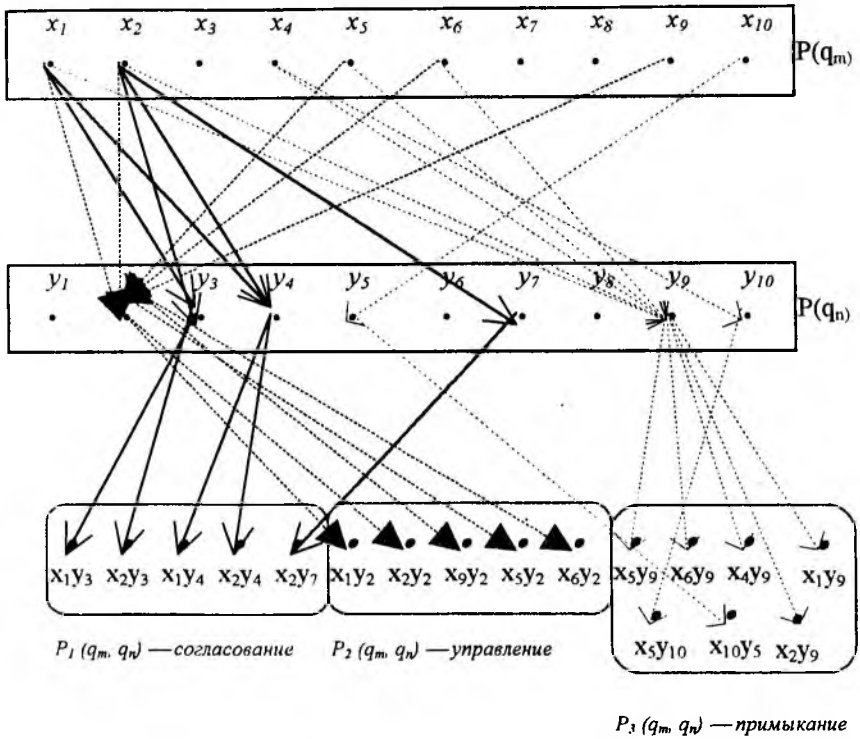
Словоформы, стоящие на первом месте в словосочетании, будут иметь следующие МИ:

x_1= *существительное, именительный падеж,*
x_2= *существительное, косвенный падеж,*
x_3= *прилагательное,*
x_4= *причастие,*
x_5= *глагол не прошедшего времени,*
x_6= *глагол прошедшего времени,*
x_7= *порядковое числительное,*
x_8= *количественное числительное,*
x_9= *предлог,*
x_{10}=*наречие.*

МИ словоформ, стоящих на втором месте в цепочке, имеют следующий вид:

y_1= *существительное, именительный падеж,*
y_2= *существительное, косвенный падеж,*
y_3= *прилагательное,*
y_4= *причастие,*
y_5= *глагол не прошедшего времени,*
y_6= *глагол прошедшего времени,*
y_7= *порядковое числительное,*
y_8= *количественное числительное,*
y_9= *предлог,*
y_{10}= *наречие.*

Связи между словоформами в словосочетаниях при трех возможных типах грамматического подчинения русского языка графически показаны на рисунке.



Аализ словосочетаний русского языка показал, что две рядом стоящие словоформы могут быть связаны в словосочетании по типу "согласование". При этом образуются следующие композиции МИ:

x_1 y_3 — существительное именительного падежа, прилагательное (стол деревянный, карандаш простой);

x_2 y_3 — существительное косвенного падежа, прилагательное (деревянном столе, простым карандашом);

x_1 y_4 — существительное именительного падежа, причастие (рисующий мальчик, начерченный план);

x_2 y_4 — существительное косвенного падежа, причастие (рисующего мальчика, начерченном плане);

x_2 y_7 — существительное косвенного падежа, порядковое числительное (третьего студента, вторым поворотом) .

Для типа связи управление характерны следующие сочетания МИ двух словоформ:

x_1y_2 — *существительное именительного падежа, существительное косвенного падежа (работа топором, ножка стула);*

$x_2\ y_2$ — *существительное косвенного падежа, существительное косвенного падежа (работой топором, ножкой стула);*

$x_9\ y_2$ — *предлог, существительное косвенного падежа (на берегу, в доме);*

x_5y_2 — *глагол не прошедшего времени, существительное косвенного падежа (вижу текст, производить погрузку) ;*

$x_6\ y_2$ — *глагол прошедшего времени, существительное косвенного падежа (произвел погрузку, увидел дом) .*

Для примыкания характерны следующие сочетания МИ рядом стоящих словоформ:

x_5y_9 — *глагол не прошедшего времени, предлог (идет на , поступает в);*

x_6y_9 — *глагол прошедшего времени, предлог (побежал на, поступил в);*

x_4y_9 — *причастие, предлог (игравший в, написавший на);*

x_1y_9 — *существительное именительного падежа, предлог (студент из);*

x_2y_9 — *существительное косвенного падежа, предлог (студентах из);*

x_5y_{10} — *глагол не прошедшего времени, наречие (одевается красиво, работает хорошо);*

$x_{10}\ y_5$ — *наречие, глагол не прошедшего времени (хорошо говорит, легко несет) .*

Для математического описания связей между МИ рядом стоящих словоформ предложения воспользуемся формулой *(2). М*ножество возможных МИ первых словоформ словосочетания задается предикатом $P(q_m)$, который может быть представлен следующим образом:

$$P(q_m) = q_m^{x1} \vee q_m^{x2} \vee q_m^{x3} \vee q_m^{x4} \vee q_m^{x5} \vee q_m^{x6} \vee q_m^{x7} \vee q_m^{x8} \vee q_m^{x9} \vee q_m^{x10}.$$

Множество МИ стоящих на втором месте словоформ может выражаться предикатом:

$$P(q_n) = q_n^{y1} \vee q_n^{y2} \vee q_n^{y3} \vee q_n^{y4} \vee q_n^{y5} \vee q_n^{y6} \vee q_n^{y7} \vee q_n^{y8} \vee q_n^{y9} \vee q_n^{y10}.$$

При грамматическом подчинении по типу согласования $\gamma_1(q_n,\ q_m)$ может быть представлена следующим образом:

$$\gamma_1(q_n, q_m) = q_n^{x1} q_m^{y3} \vee q_n^{x2} q_m^{y3} \vee q_n^{x1} q_m^{y4} \vee q_n^{x2} q_m^{y4} \vee q_n^{x2} q_m^{y7};$$

при грамматическом подчинении по типу управления $\gamma_2(q_n,\ q_n)$ представляется формулой

$$\gamma_2(q_n, q_m) = q_n^{x1} q_m^{y2} \vee q_n^{x2} q_m^{y2} \vee q_n^{x9} q_m^{y2} \vee q_n^{x5} q_m^{y2} \vee q_n^{x6} q_m^{y2};$$

при примыкании $\gamma_3(q_n, q_m)$ определяется как

$$\gamma_3(q_n, q_m) = q_n^{x5} q_m^{y9} V q_n^{x6} q_m^{y9} V q_n^{x4} q_m^{y9} V q_n^{x1} q_m^{y9} V q_n^{x2} q_m^{y9} V q_n^{x5} q_m^{y10} V q_n^{x10} q_m^{y5}.$$

Тогда в соответствии с формулой *(2)* опишем множество возможных связей комплексов морфологической информации в словосочетаниях по типу согласования, задаваемое с помощью предиката $P_1(q_m, q_n)$:

$$P_1(q_m, q_n) = \gamma_1(q_m, q_n) \bullet P(q_m) \bullet P(q_n) = (q_n^{y1} q_m^{y3} V q_n^{y2} q_m^{y3} V q_n^{y1} q_m^{y4} V q_n^{y2} q_m^{y4} V q_n^{y2} q_m^{y7}) (q_m^{x1} V q_m^{x2} V q_m^{x3} V q_m^{x4} V q_m^{x5} V q_m^{x6} V q_m^{x7} V q_m^{x8} V q_m^{x9} V q_m^{x10})(q_n^{y1} V q_n^{y2} V q_n^{y3} V q_n^{y4} V q_n^{y5} V V q_n^{y6} V q_n^{y7} V q_n^{y8} V q_n^{y9} V q_n^{y10}). \quad (3)$$

Множество возможных связей комплексов морфологической информации в словосочетаниях по типу управления, задаваемое с помощью предиката $P_2(q_m, q_n)$:

$$P_2(q_m, q_n) = \gamma_2(q_m, q_n) \bullet P(q_m) \bullet P(q_n) = (q_n^{x1} q_m^{y2} V q_n^{x2} q_m^{y2} V q_n^{x9} q_m^{y2} V q_n^{x5} q_m^{y2} V V q_n^{x6} q_m^{y2}) (q_m^{x1} V q_m^{x2} V q_m^{x3} V q_m^{x4} V V q_m^{x5} V q_m^{x6} V q_m^{x7} V q_m^{x8} V q_m^{x9} V q_m^{x10}) (q_n^{y1} V q_n^{y2} V q_n^{y3} V q_n^{y4} V q_n^{y5} V q_n^{y6} V V q_n^{y7} V q_n^{y8} V q_n^{y9} V q_n^{y10}). \quad (4)$$

Множество возможных связей комплексов морфологической информации в словосочетаниях по типу примыкания, задаваемое с помощью предиката $P_3(q_m, q_n)$, можно описать следующим образом:

$$P_3(q_m, q_n) = \gamma_3(q_m, q_n) \bullet P(q_m) \bullet P(q_n) = (q_n^{x5} q_m^{y9} V q_n^{x6} q_m^{y9} V q_n^{x4} q_m^{y9} V q_n^{x1} q_m^{y9} V q_n^{x2} q_m^{y9} V q_n^{x5} q_m^{y10} V V q_n^{x10} q_m^{y5}) (q_m^{x1} V q_m^{x2} V q_m^{x3} V q_m^{x4} V q_m^{x5} V q_m^{x6} V q_m^{x7} V q_m^{x8} V q_m^{x9} V V q_m^{x10}) (q_n^{y1} V q_n^{y2} V q_n^{y3} V q_n^{y4} V q_n^{y5} V q_n^{y6} V q_n^{y7} V q_n^{y8} V q_n^{y9} V q_n^{y10}). \quad (5)$$

При подстановке КМИ первой и второй словоформы словосочетания, полученных на этапе морфологического анализа, в формулы (3)—(5) предикаты, которые описывали тип словосочетания, не присущий данным словоформам, обращаются в нуль. Те же предикаты, которые принимают значение 1, позволяют существенно снизить возможные варианты сочетаний между словоформами. Таким образом, полученные бинарные предикаты (формулы (3)—(5)) позволяют уже на этапе предсинтаксиса снять часть морфологической омонимии, что способствует улучшению качества перевода, особенно при переводе с флективных (русского, украинского) языков.

Список литературы: 1. *Белоногов Г.Г., Кузнецов Б.А.* Языковые средства автоматизированных информационных систем. М.: Наука, 1983. 288 с. 2. *Хайрова Н.Ф., Замаруева И.В.* Машинный перевод. Учеб.пособ. Х.:Око, 1998. 82 с. 3. *Шабанов-Кушнаренко Ю.П.* Теория интеллекта. Математические средства. Х.: Вища шк.Изд-во при Харьк.ун-те, 1984. 144 с. 4. *Шабанов-Кушнаренко Ю.П., Шаронова Н.В.* Компараторная идентификация лингвистических объектов.К.:ИСИО, 1993. 116 с.

Поступила в редколегию 20.10.98