Эксперимент по автоматизации определения семантики валентных связей с помощью машинного обучения

Вячеслав Диконов
ИППИ РАН
dikonov@iitp.ru

Резюме

В статье описывается ход и первые результаты эксперимента, в ходе которого алгоритм машинного обучения на основе SVM применяется для определения того, каким семантическим ролям соответствуют зафиксированные в комбинаторном словаре ЭТАП синтаксические валентности. Результаты применения описанного метода могут быть полезны для семантического анализа текста и разработки онтологии предикатов[1]. К моменту написания данной статьи эксперимент не завершен, но получены первые обнадеживающие результаты.

1. Введение

Одним из важных направлений работы в лаборатории компьютерной лингвистики ИППИ РАН является автоматический и полуавтоматический семантический анализ текста, запись его смысла в формальном виде и извлечение фактов с помощью системы ЭТАП [1]. При анализе текста чрезвычайно важную роль играют предикатные слова, которые занимают центральное место в синтаксической и семантической организации текста на естественном языке. Для корректного разбора на синтаксическом и семантическом уровне важно знать их валентные свойства. Они определяют, каким образом и какие слова и понятия (называемые при этом актантами или аргументами) могут присоединяться к определенному предикату, чтобы получилось грамматически правильное и осмысленное предложение.

При переходе к семантической структуре те слова, которые занимают синтаксические валентные места предикатного слова, истолковываются как семантические актанты. Это позволяет сказать, какие существенные части описываемой в предложении ситуации важные для ее осмысления они представляют. При этом возникает необходимость определить семантическую роль каждого из имеющихся у данного предиката актантов. Например:

Вася $^{предик \rightarrow \textbf{кто?}}$ *отправился в гости* $^{2\text{-компл} \rightarrow \textbf{куда?}}$.

2. Используемые ресурсы

Ключевую роль в описываемой работе играют лингвистические ресурсы: комбинаторный словарь ЭТАП (КС), словарь UNL [2] и онтология SUMO [3,4]. В них содержатся необходимые данные: описание синтаксических валентностей предикатных слов (КС), детализированные лексические значения (UNL), классификация предикатов (онтология).

Комбинаторный словарь состоит из лексем, т. е. единиц более мелких чем слово, которые теоретически должны представлять слово в одном из значений. Практически же лексемы в используемых словарях системы ЭТАП соответствуют пакетам из нескольких родственных лексических значений, которые могут иметь разные формальные определения и разные наборы семантических актантов. Разные лексемы в КС отличаются друг от друга по существенным синтаксическим и комбинаторным свойствам. КС подробно описывает синтаксические свойства слов и содержит отрывочные и обобщенные сведения об их семантике.

В системе ЭТАП также имеется словарь семантического языка UNL, где каждая статья соответствует ровно одному лексическому значению. Значения обозначаются с помощью «универсальных слов» UNL (UW). В настоящее время словарь UNL в ЭТАП содержит 1) краткие неформальные определения и примеры на английском языке, 2) ссылки на онтологию SUMO и Wordnet [5] 3) переводы на английский и

[1] Эксперимент проводится в рамках работы поддержанной грантом РФФИ 11-06-00405 “Разработка онтологии для автоматической обработки текстов”

русский и 4) правила для перехода от синтаксических связей к семантическим отношениям UNL. Последние и составляют предмет описываемого эксперимента.

Валентности и семантические роли

Каждая статья словаря КС содержит модель управления (МУ), которая описывает возможные способы подсоединения актантов в синтаксическом дереве. Она представляет собой таблицу, где колонки соответствуют возможным для данного предиката аргументным синтаксическим отношениям (*1-, 2- ,3-, 4- и 5-комплетивные, предикативное и квазиагентивное)*. Эти связи неповторимы, то есть при одном предикате не может быть нескольких аргументных связей одинакового типа. Строки таблицы МУ содержат перечень лексических и синтаксических средств и ограничений для заполнения каждой валентности в виде предлогов, падежей, синтаксических признаков и иногда семантических дескрипторов, которыми должен обладать актант (Рис.1). Информация из МУ используется в ЭТАП для построения синтаксических структур.

predic	1-compl	2-compl	3-compl	4-compl
'ЛИЦО', 'ТРАНСПОРТ' (сем. дескрип-торы)	ILLAT (предлоги со значением движения ОТ чего-то)	ALLAT (предлоги со значением движения К чему-то)	FOR1	BY1
	S (существи-тельное)			

Рис. 1 Модель управления слова DEPART

Правила в статьях словаря UNL заменяют синтаксические связи на семантические ролевые отношения, которые определяются индивидуально для каждого лексического значения и являются общими для задаваемых онтологией классов предикатов. Примеры семантических ролей в нотации UNL: ***agt*** *(агент),* ***obj*** *(объект),* ***ben*** *(бенефициар),* ***rec*** *(реципиент),* ***tim*** *(время),* ***plc*** *(место),* ***pur*** *(цель)* и др. В большинстве случаев правила описывают простую таблицу соответствий между синтаксическими связями и семантическими ролями (Рис.2).

DEPART
predic = agt (агент, кто?)
1-compl = plf (исходное место, откуда?)
2-compl = plt (конечное место, куда?)
3-compl = pur (цель, зачем?)
4-compl = met (транспорт, на чем?)

Рис.2 Соответствие синтаксических связей и семантических ролей для слова DEPART в значении «отправиться»

Такое идеальное соответствие может нарушаться. Одной лексеме КС системы ЭТАП часто соответствуют несколько UW с различными наборами семантических актантов, а имеющаяся МУ этого не учитывает. Это порождает случаи, когда синтаксическая валентность не имеет смысла в конкретном значении слова и не может быть связана ни с одной семантической в рамках данного значения, либо соответствующая семантическая валентность не является одной из ключевых частей определения значения и не может считаться настоящим семантическим актантом (Рис.3).

APPEAR *в значении «казаться»*
predic = aoj (кто/что?)
1-compl [TO2, A] = obj (каким?)
2-compl [FROM]= **ничего**
3-compl [TO1] = ben (кому?)

MIZZLE *в значении «моросить»*
predic = obj (что?)
1-compl [ILLAT] = **plf** (откуда?) **не актант**
2-compl [ALLAT]= **plt** (куда?) **не актант**
Две последние валентности не являются обязательными для толкования «моросить».

REPRESENT *в значении «представлять что кому»*
predic = aoj (кто/что?)
1-compl [S, THAT1] = obj (что?)
2-compl [AS1,TO1]= **cob/rec (как что/кому?)**
3-compl [LOCAT...] = scn (где?)
Некорректная МУ лексемы, представляющей два лексических значения. Одна колонка соответствует двум семантическим ролям.

Рис.3 Различные аномалии в таблице соответствия актантов и ролей

Также бывает, что одна синтаксическая валентность соответствует двум разным ролям, или семантическая роль не имеет соответствия в МУ. Последние два случая указывают на возможную ошибку в написании статьи КС.

3. Постановка задачи

Существует проблема неполноты лингвистической информации в системе ЭТАП, когда при семантическом анализе оказывается, что семантика присоединяющего конкретный актант отношения точно неизвестна. Это указывает на пробелы в словаре. Во время ручного редактирования семантического словаря UNL стало ясно, что а) трудоемкость этой нужной работы в масштабах всего словаря очень высока, б) существует заметная корреляция между семан-

тикой валентных мест предикатов и используемыми языком лексическими и синтаксическими средствами для их заполнения. Этот факт позволил сформулировать эвристические правила для автоматического задания некоторых соответствий между синтаксическими связями и типичными ролями. Однако, этот простой подход не позволяет решить задачу полностью. Для успешной категоризации оставшихся случаев необходимо выявить и учесть все большее число разнообразных сочетаний известных факторов. Вместе с тем, задачи такого рода успешно решаются с помощью алгоритмов машинного обучения, а имеющаяся лингвистическая информация позволяет представить проблему в пригодном для их применения виде.

Таким образом, цель проводимого эксперимента — выяснить, возможно ли повысить качество интерпретации валентностей путем дополнения правил соответствия синтаксических валентностей и семантических ролей с помощью стандартных алгоритмов машинного обучения на основе SVM в масштабе всего словаря. Хороший результат будет означать высокую корреляцию между семантическими ролями предикатов определенного класса и поверхностными способами их выражения в естественном языке. В случае успеха предполагается сопоставить полученную информацию с данными логического вывода онтологии о семантических ролях предикатов. Сопоставление может выявить ошибки и упущения в собственно онтологии, в комбинаторном словаре или в отнесении предикатов к определенным классам онтологии.

4. Исходный материал

До начала эксперимента уже была проделана значительная работа по ручному редактированию наборов семантических ролей и их соответствий с аргументными синтаксическими связями. В основном она затронула значения наиболее частотных английских глаголов. Для некоторых валентностей оставшихся предикатов соответствия были установлены при помощи правил.

После этого был проведен анализ словарей системы ЭТАП, чтобы определить, какие статьи словаря UNL содержат неполный набор соответствий между валентностями МУ и семантическими ролями UNL. Из множества найденных UW была выделена группа из 4800 приоритетных значений, которые требовалось отредактировать в первую очередь. В нее вошли лексические значения английских глаголов, которые а) имеют неполную таблицу соответствий связей и ролей и б) являются наиболее частотными среди возможных для каждого глагола (по данным открытого корпуса Semcor), либо могут быть выбраны при автоматическом анализе текста с помощью уже имеющихся правил разрешения лексической неоднозначности.

Для дальнейшего редактирования приоритетной группы UW была построена специальная таблица на основе данных словарей КС и UNL. На рисунке 4 приводится одна строка этой таблицы (первый столбец на рисунке) и комментарий к ее содержимому (второй столбец).

25	*частотность*
RECOGNIZE1	*лексема КС/слово*
recognize(icl>do,agt>thing,obj>thing)	*UW*
"Detect with the senses"	*глосс*
WORDNET	*источник/редактор*
1-COMPL	***синт. валентность 1***
S	*существительное*
OBJ	***что?***
2-COMPL	***синт. валентность 2***
BY1	*управление через предлог by*
MET	***каким методом?***
PREDIC	***синт. валентность 0***
S	*существительное*
AGT>living_thing	***кто? (живое существо)***
02193194;3	*ссылка на Wordnet*
Perception	*ссылка на SUMO*

Рис.4: Одна строка таблицы для ручного редактирования.

Она является частью намного большей полной таблицы, включающей менее приоритетный словарный материал, который тоже требует аналогичного редактирования.[2] Данные из отредактированных строк таблицы впоследствии могут импортироваться в словари ЭТАП в виде искомых правил соответствия связей и ролей.

5. Ход эксперимента

Из 4800 включенных в таблицу UW вручную были отредактированы примерно 3000. Для заполнения оставшейся части таблицы был

[2] 20753 лексемы английского КС снабжены моделями управления. Соответствующие им UW могут быть материалом для разметки.

создан специализированный статистический таггер, который основывается на svm-tools [6,7] и библиотеке libsvm. Отредактированные вручную строки составили используемую обучающую выборку, а оставшаяся часть таблицы используется для тестирования.

Обучение таггера

Общий принцип устройства статистического таггера таков, что рассматриваемую проблему необходимо сначала преобразовать в удобный для машинного обучения вид. В данном эксперименте это означает формирование серии специальных таблиц — по одной на каждую из семантических ролей, а также два особых флага: невозможности заполнения синтаксической валентности в данном лексическом значении и периферийности соответствующей семантической роли (см. Рис.3 выше). Всего получено 34 таблицы. Каждая из них представляет собой список всех имеющихся в обучающей выборке валентностей без привязки к UW. Каждая валентность каждого из предикатов рассматривается отдельно. Первая колонка обучающих таблиц содержит условный номер соответствующей семантической роли (1...34) или 0. Остальные колонки содержат перечень признаков, на основании которых таггер должен классифицировать валентности.

Набор признаков состоит из двух групп: синтаксической и семантической. Во время написания статьи первая состоит из:

- имени синтаксической связи;
- соответствующих этой связи строк МУ, соединенных вместе;
- отдельных частей содержимого МУ: предлогов, синтаксических признаков ЭТАП, дескрипторов,

а вторая включает в себя:

- набор уже известных (ранее приписанных вручную или правилами) семантических ролей для данного значения;
- предписанный стандартом UNL класс для значений глаголов (действия, состояния, спонтанные события);
- перечень классов онтологии, в которые входит данный предикат;
- полный путь между вершинным классом Thing и терминальным классом данного предиката.

Все признаки шифруются условными номерами. Значения признаков бинарны.

В режиме обучения таггер формирует 34 модели, которые позволяют классифицировать отдельные валентности предикатов и определять вероятность того, что при данном наборе признаков следует приписать соответствующую модели семантическую роль или один из двух особых флагов.

Для формирования статистических моделей используется нелинейное svm-ядро.

Разметка

В режиме разметки программа читает размечаемый материал в формате описанной выше (Рис.4) таблицы для ручного редактирования и формирует 34 таблицы, аналогичных обучающим таблицам. При этом используется специальный словарь соответствия имен признаков условным цифровым кодам. Затем SVM-таггер применяет ранее полученные модели. Каждая модель классифицирует размечаемые наборы признаков синтаксических валентностей и предиката в целом на два класса и сообщает меру вероятности положительного и отрицательного решения. После применения всех моделей образуется ранжированный по степени вероятности список семантических ролей, которые могут быть приписаны конкретной синтаксической валентности.

Следует отметить, что таггер рассматривает не предикат в целом, а каждую валентность отдельно от предиката в виде набора признаков. Взаимозависимости между парными семантическими ролями, такими как объект и ко-объект (obj и cob), начальное состояние и конечное состояние (src и gol) и т. п. содержатся имплицитно в наборе признаков и должны быть определены самим таггером. Впоследствии проводится контроль таких пар с помощью правил.

После применения моделей программа восстанавливает соответствия между размеченными наборами признаков и ячейками таблицы для ручного редактирования и заполняет пустые ячейки и ячейки, которые ранее были заполнены эвристическими правилами и помечены как сомнительные.

6. Предварительные результаты

На данный момент получены только предварительные результаты. Svm-tools имеют режим самоконтроля, в котором предварительно размеченная обучающая выборка делится на несколько частей. Одна часть используется для контроля, а все остальные — для обучения. Затем процесс повторяется с другим разбиением на

части. В результате такого самотестирования были получены очень высокие результаты: от 96 до более чем 99% для всех классов кроме двух. Для класса agt (агент) результат составил примерно 80%, а для класса obj (объект) — 68%. При этом следует отметить, что по правилам UNL эти две метки используются в синтаксически идентичных контекстах, когда они соответствуют одной и той же предикативной синтаксической валентности**.** Кроме того, метка obj используется шире прочих, так как фактически объединяет в себе роли пациенса, содержимого сообщения и некоторые другие. Разнообразие наборов признаков для obj и число употреблений больше чем для других меток ролей. Однако, обе эти роли очень хорошо коррелируют с синтаксическими отношениями «predic» и «1-compl». Это позволяет контролировать их использование с помощью дополнительных внешних правил.

7. Существующие проблемы

К сожалению, несмотря на формально хороший процент точности разметки есть и проблемы. Главной проблемой оказалась плохая переносимость моделей на новый материал. Когда таггер был применен для разметки ранее никогда не использовавшегося материала с более низким приоритетом, то по результатам ручной проверки его эффективность резко упала. Максимальный полученный результат составил 75%, а минимальный — 60%.

Также выяснилось, что во многих случаях таггер оказывался неспособен приписать какую-либо роль, и доля таких случаев в новом материале достигла примерно 50%. Кроме того, в размеченных таггером валентностях МУ часто происходит взаимная путаница между семантически близкими ролями, например ins «инструмент» и met «метод осуществления», src «исходное состояние» и plf «исходное место», то есть существующий таггер не мог их уверенно различать.

Основным объяснением этого явления представляется все еще крайне малый по меркам статистики объем обучающей выборки. Около 3000 вручную размеченных UW дают 6381 набор признаков без учета повторений (по одному набору на каждую валентность каждого UW). Это число очень неравномерно распределяется на 33 класса. Особый признак семантической периферийности синтаксического актанта является модификатором, который добавляется к прочим ролям. Неравномерность распределения примеров по разным классам связана с частотностью употребления разных семантических ролей. Некоторые из них представлены буквально парой десятков примеров.

Кроме того, важнейшим критерием для классификации валентностей предиката является его онтологический класс. Используемая онтология содержит свыше 19000 классов и далеко не все они были представлены в обучающей выборке. Таггер не может использовать неизвестные ему признаки онтологических классов. Если размечаемый предикат относится к одному из таких неизвестных классов, и вышележащих классов недостаточно для успешной классификации, то единственными действенными критериями становятся признаки синтаксической группы. Результат работы таггера при этом оказывается ненадежным.

Перечисленные проблемы кажутся разрешимыми. Результаты разметки новых фрагментов таблицы могут быть улучшены за счет расширения обучающего набора примеров, более детальной оценки полезности отдельных признаков для решения задачи и добавления внешнего механизма контроля на основе онтологической информации и правил.

Литература

[1] Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин и др. *Лингвистический процессор для сложных информационных систем*. М.: Наука, 1992

[2] Boguslavsky I.M., Dikonov V.G. *Universal Dictionary of Concepts* MONDILEX First Open Workshop “Lexicographic Tools and Techniques“. Moscow, Russia. October 3-4, 2008. C. 31-41. ISBN: 978-5-9900813-6-9.

[3] Adam Pease *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA. ISBN 978-1-889455-10-5.

[4] http://www.ontologyportal.org/

[5] Christiane Fellbaum *WordNet: An Electronic Lexical Database.* Bradford Booksб 1998

[6] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM : a library for support vector machines*. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm