Подходы к решению проблемы определения границ предложения в английском языке

Безвербный И.Г.

Омская государственная медицинская академия,

ogma@omsk.net.ru

1. **Введение.**

Определение границ предложения является важной задачей для многих приложений компьютерной обработки текстовых массивов, таких как синтаксических анализаторов, таггеров, систем машинного перевода и др. (Gale and Church, 1993), (Kay and Roscheinsen, 1993). Для успешной работы таких программ необходимо четкая маркировка окончаний предложений, что обычно достигается путем подстановки определенных символов в конец каждого предложения, таким образом чтобы программа анализирующая текст могла бы легко выделить предложение в тексте (Palmer and Hearst, 1994).

В описаниях свободно распространяемых приложений для компьютерной обработки текстов указывается, что до того как программа начинает анализировать данные, часто существует необходимость подразделения текста на предложения, но не указывается каким образом этого можно достигнуть. Другие авторы подразумевают такое подразделение, но не обсуждают его выполнение (Cutting et al., 1992).

На первый взгляд может показаться, что поиск и маркировка в тексте знаков пунктуации, которые обычно обозначают конец предложения, таких как ".", "?" и "!" будет достаточным для успешного выполнения поставленной задачи. Однако все вышеперечисленные знаки являются неоднозначными, так как используются не только для обозначения границ предложений. Например, точка, в английском и американском варианте орфографии, может обозначать десятичную дробь, сокращение, сокращение в конце предложения, разделение доменов в адресах электронной почты и т.д. Восклицательный и вопросительный знаки, хотя и являются менее неоднозначными, так же могут, наряду с окончанием предложения, встречаться в прямой речи, для выделения или усиления какого либо высказывания. Неоднозначность использования этих знаков пунктуации можно показать на следующих примерах (Palmer, 1994):

- The group included Dr. J.M. Freeman and T. Boone Pickens Jr.
- It was due Friday by 5 p.m. Saturday would be too late.
- She has an appointment at 5 p.m. Saturday to get her car fixed.

Существование пунктуации в грамматических подпредложениях предполагает возможность дальнейшего подразделения границ предложений на типы (например когда знаки окончания предложения встречаются в кавычках в прямой речи):

- "This issue crosses party lines and crosses philosophical lines!" said Rep. John Rowland (R., Conn.).

2. Подходы к решению проблемы.

Прежде, чем приступить к описанию и оценке алгоритмов определения границ предложения необходимо определить базовый алгоритм с которым можно проводить сравнение различных подходов к решению проблемы. Большинство авторов за основу принимают простейший алгоритм поиска всех знаков, которыми обозначают конец предложения. такой алгоритм позволяет выявлять границы предложения с точностью, которая является самой низкой.

Таким образом, хороший алгоритм будет иметь точность намного превышающую точность базового (Palmer, 1994).

Насколько известно в за последнее время было несколько публикаций посвященных проблеме определения границ предложений. Большинство из этих работ используют в качестве алгоритма поиска регулярную грамматику, обычно совместно с ограниченным просмотром слов находящихся перед и за знаком границы предложения, т.е. в упрощенном виде такая грамматика предполагает поиск шаблона "точка – пробел - прописная буква", который обычно находится в конце предложения. Более сложные системы рассматривают целое слово которое предшествует или следует за знаком пунктуации и сравнивают их с обширным списком сокращений и имен собственных.

Christiane Hoffman(1994) использовала регулярную грамматику для классификации знаков препинания при исследовании сборника текстов немецкой газеты "DieTageszeitung" с точностью базового алгоритма 92% при этом она использовала обширный словарь сокращений для определения частотности точек в соответствии с их наиболее вероятной функции в предложении. Ее метод позволил выявить границы свыше 98% предложений. Однако ее алгоритм был специально разработан для анализа текста газеты "Die Tageszeitung", и, как указывает автор, точность распознания предложений в других текстовых массивов будет зависеть от качества словаря сокращений применяемых в данном тексте.

Mark Wasson разработал систему распознания границ предложений, в словарь которой он включил 18002 слова (сокращения, термины отсутствующие в словаре и т.д.) и 1419 положений знаков препинания в тексте. Точность распознания составила 99,7%. В то же время, необходимо отметить. что его словарь, главным образом, составлялся на основе текстов относящихся к судопроизводству, поэтому, как он сам отмечает, существует вероятность, что его алгоритм будет работать намного хуже при анализе текстов другой тематики.

Riley (1989) создал алгоритм, который использует подход так называемых ниспадающих деревьев (Breiman et al., 1984) для определения границ предложений в соответствии со следующими параметрами:

- Вероятность [слово предшествующее "." встречается в конце предложения]
- Вероятность [слово следующее за "." встречается в начале предложения]
- Длина слова предшествующего "."
- Длина слова следующего за "."
- Регистр слова слова предшествующего ".": верхний, нижний, цифры
- Регистр слова слова следующего за ".": верхний, нижний, цифры
- Знаки препинания после "." (если есть)
- Является ли слово предшествующее "." сокращением

Алгоритм использует информацию о словах которые находятся с обоих сторон от знака точки и сверяясь со словарем делает вывод о вероятности границы предложения. Процент вероятности был получен путем обработки заранее отмеченных 25 млн. слов взятых из сборника новостей информационного агентства "Associated Press ". Алгоритм был протестирован при выполнении анализа сборника текстов Brown University, точность составила 99,8%.

David Palmer(1994) разработал систему SATZ, алгоритм которой основан на использовании словаря в котором указана вероятность того, что та или иная часть речи может встречаться в конце предложения, и обучаемой сети для быстрой адаптации к тому или иному тексту. Хотя, точность распознания границ предложения, как указывает автор, превышает 99%, недостаток состоит в необходимости перед каждым анализом обучать программу анализируя заранее размеченный текст.

3. **Решение проблемы с помощью программы DELIMITATOR.**

Все вышеперечисленные подходы к решению проблемы разграничения предложений, хотя и имеют очень высокую точность распознавания, но являются контекстно-зависимыми, т.е. все системы разрабатывались для определенных текстовых массивов и применение их для обработки текстов принадлежащих к другим областям знаний или деятельности оказывается неудачным, и все они требуют предварительной обработки текстов.

Принимая все это во внимание, мы поставили перед собой задачу разработки компьютерной программы для автоматического определения границ предложения, которая бы отвечала следующим требованиям:

1. Система должна быть контекстно-независимой.
2. Система поиска должна быть полностью автоматизированной.
3. Система не должна содержать большой словарь.
4. Анализ, проводимый программой должен быть как можно более точен.
5. Анализ не должен занимать много времени.

Взяв за основу эти требования и проанализировав доступные системы определения границ предложения, мы пришли к выводу, что наиболее реальный контекстно-независимый алгоритм для сегментации предложений должен быть основан на регулярной грамматике и правилах исключений.

В настоящее время алгоритм применяется в программе DELIMITATOR.

Программа обрабатывает текстовый массив следующим образом: возможные границы предложений распознаются путем поиска в обрабатываемом текстовом массиве последовательностей символов разделенных промежутком и содержащих символы, обычно обозначающие конец предложения. После нахождения такой последовательности, программа анализирует символы, находящиеся справа и слева от найденных возможных границ предложений, чтобы исключить возможную ошибку в определении границы, например, в случае нахождения аббревиатуры с точкой в конце.

После этого программа составляет карту или формулу предложения, в которой всем словам в массиве присваивается определенный индекс, показывающий к какой части речи, возможно, относится слово и на основе этой индексации проводится дополнительная обработка текста для устранения ошибок при определении границы предложения (например: если слово стоящее перед точкой имеет индекс артикля, то можно с большей степенью уверенности считать, что данная точка не является знаком, обозначающим окончание предложения и т.п.).

К сожалению, так как работа над проектом создания данного алгоритма находится в самом начале, то средняя точность определения границ составляет 86,4% при нижней границе 53% (на начальном этапе разработки эта пропорция составляла 74% к 53%).

Мы продолжаем улучшать алгоритм, вводя новые правила и исключения и можно надеяться, что его точность приблизится к максимально возможному пределу, и будет сравнима с уже существующими лучшими системами, при неоспоримом преимуществе контекстной независимости.

Литература

Антрушина Г.Б. и др. Лексикология английского языка. М., 1999.

Штелинг Д.А. Грамматическая семантика английского языка. М., 1996

Gale W.A., Church K.W. A program for aligning sentences in bilingual corpora. Computational Linguistics. 1993.

Kay M., Roscheinsen. Text-translation alignment. Computational Linguistics. 1993.

Palmer D.D., Hearst M.A. Adaptive sentence boundary disambiguation. Stuttgart. 1994.

Cutting D. et al. A practical part-of-speech tagger. Trento. 1991.

Breiman L. et al. Classification and regression trees. Belmont. 1984.

Kavanagh J. The text analyzer: A tool for extracting knowledge from text. University of Ottawa. 1999.