Антонов Егор Сергеевич

АПРИОРНАЯ ИНФОРМАЦИЯ КАК СПОСОБ РАЗРЕШЕНИЯ ОНТОЛОГИЧЕСКОЙ И ЯЗЫКОВОЙ ОМОНИМИИ

В статье рассматривается целесообразность использования априорной информации для разрешения языковой и онтологической омонимии именованных сущностей. На материале размеченного корпуса из 1700 англо-язычных новостных статей опробована стратегия выбора наиболее вероятного объекта с двумя настраиваемыми параметрами (минимальная вероятность соответствия, минимальное количество упоминаний в корпусе). Подобная стратегия позволяет достигнуть большой точности разрешения омонимии, однако ее применение не имеет смысла при большом количестве объектов онтологии из-за низкой полноты.

Адрес статьи: www.gramota.net/materials/2/2012/6/2.html

Источник

Филологические науки. Вопросы теории и практики

Тамбов: Грамота, 2012. № 6 (17). С. 15-19. ISSN 1997-2911.

Адрес журнала: www.gramota.net/editions/2.html

Содержание данного номера журнала: www.gramota.net/materials/2/2012/6/

3. **Ковалева Н. А.** Авторское фразообразование и коммуникативная стратегия текста в письмах А. П. Чехова. Астрахань: Изд-во Астраханского ун-та, 2000. 247 с.
4. **Кунин А. В.** Фразеологическая вариантность и структурная синонимия в современном английском языке // Проблемы фразеологии и задачи ее изучения в высшей и средней школе. Вологда: Северо-Западное книжное издательство, 1967. С. 146-153.
5. **Лаптева О. А.** Узус // Русский язык: энциклопедия / гл. ред. Ю. Н. Караулов. М.: Дрофа, 1998. С. 583-584.
6. **Межжерина С. А.** Взаимодействие фразеологического оборота и контекста в художественной речи // Русский язык в школе. 1971. № 3. С. 75-78.
7. **Пекарская И. В.** Контаминация в контексте проблемы системности стилистических ресурсов русского языка. Абакан: Издательство Хакасского государственного университета им. Н. Ф. Катанова, 2000. Ч. I. 248 с.
8. **Федоров А. И.** Образная речь. Новосибирск: Наука, 1985. 120 с.
9. **Фразеологический словарь русского литературного языка** / сост. А. И. Федоров. М.: Цитадель, 1997. Т. 1.
10. **Фразеологический словарь русского литературного языка** / сост. А. И. Федоров. М.: Цитадель, 1997. Т. 2.
11. **Фразеологический словарь русского языка** / под ред. А. И. Молоткова. М.: Советская энциклопедия, 1967. 544 с.
12. **Шанский Н. М.** Основные свойства и приемы стилистического использования фразеологических оборотов // Русский язык в школе. 1957. № 3. С. 13-21.
13. **Шмелев Д. Н.** Современный русский язык. Лексика. М.: Просвещение, 1977. 334 с.

PHRASEOLOGICAL UNIT IN LITERARY DISCOURSE: TO PROBLEM OF ELOCUTIONARY STATUS

Yuliya Valer'evna Aleshechkina
Irina Vladimirovna Pekarskaya, Doctor in Philology, Professor
Department of Russian Language Stylistics and Journalism
Khakass State University named after N. F. Katanov
aleshechkinayuliya@mail.ru

The authors consider phraseological unit in its language and speech transformations in terms of its expression, pay special attention to determining elocutionary status as an ornamental one, namely as elocutionary figurative means, and formulate the definitions of usual, transformed and occasional phraseological units.

Key words and phrases: phraseological unit; elocutionary figurative means; usual phraseological unit; individual-authorial transformations of phraseological unit; individual-authorial (occasional) phraseological unit.

УДК 81’322
Филологические науки

В статье рассматривается целесообразность использования априорной информации для разрешения языковой и онтологической омонимии именованных сущностей. На материале размеченного корпуса из 1700 англоязычных новостных статей опробована стратегия выбора наиболее вероятного объекта с двумя настраиваемыми параметрами (минимальная вероятность соответствия, минимальное количество упоминаний в корпусе). Подобная стратегия позволяет достигнуть большой точности разрешения омонимии, однако ее применение не имеет смысла при большом количестве объектов онтологии из-за низкой полноты.

Ключевые слова и фразы: распознавание именованных сущностей; разрешение омонимии именованных сущностей; онтология; априорная информация; географические объекты; новостные тексты.

Егор Сергеевич Антонов
Кафедра теоретической и прикладной лингвистики
Московский государственный университет им. М. В. Ломоносова
me@eantonov.name

АПРИОРНАЯ ИНФОРМАЦИЯ КАК СПОСОБ РАЗРЕШЕНИЯ ОНТОЛОГИЧЕСКОЙ И ЯЗЫКОВОЙ ОМОНИМИИ©

Современные базы данных об объектах действительности содержат миллионы записей. Так, в англоязычной версии Википедии содержится более 4 млн статей [6], в онтологии *Freebase* – 23 млн сущностей [5] и т.д. Наличие столь больших объемов информации позволяет людям узнать о любом интересующем их объекте, однако поиск этой информации затрудняется из-за различных проблем, в т.ч. из-за проблем омонимии. С одной стороны, имя объекта действительности может совпадать с общеупотребительным словом (**языковая омонимия**). С другой стороны, одному и тому же имени может соответствовать много объектов действительности (**онтологическая омонимия**). В качестве примера можно привести имя «Образование», которое, во-первых, является общеупотребительным словом, а во-вторых, таким именем называются сразу несколько объектов (банк, федеральная целевая программа, журнал).

Материал исследования

Первым шагом к определению связи между онтологическим объектом и текстом является определение гипотез именованных сущностей (ИС) в тексте (этап распознавания сущностей). Некоторые классы объектов

онтологии (например, персоны) могут иметь довольно сложные шаблоны распознавания. Чтобы избежать проблем с этапом определения гипотез ИС, в качестве исходного материала были выбраны географические объекты онтологии *Freebase*. На основе данных о заголовке топика из *Freebase* и идентификаторов Википедии для каждого объекта были выделены имена, под которыми тот может упоминаться в тексте (чуть более 4 млн имен для 1 млн объектов). Таким образом, этап распознавания гипотез ИС сводится к простому поиску 4 млн подстрок. Эта задача была решена с помощью алгоритма Рабина-Карпа [1]. На основе 1700 англоязычных новостных статей был размечен тестовый корпус из 32 тыс. гипотез ИС. Полученные гипотезы нуждаются в последующей обработке. Во-первых, необходимо разрешение коллизий имен (случаи, когда позиции нескольких имен пересекаются в тексте), во-вторых, нужно отсеять «ложные срабатывания», т.е. решить проблему языковой омонимии, в-третьих, нужно из списка объектов базы данных (БД) с данным именем выбрать наиболее подходящий (решение проблемы онтологической омонимии). Первый тип пост-обработки принято называть разрешением структурной омонимии [2]. В нашем корпусе масштаб структурной омонимии оказался незначительным, и большинство случаев удалось решить с помощью простых эвристических правил (удаление вложенных позиций, выбор наиболее длинного из пересекающихся имен). Остальные виды пост-обработки оказывают более значимое влияние на итоговый результат и решаются более трудоемкими методами.

Рост омонимии при больших размерах базы сущностей

Проблемы языковой и онтологической омонимии слабо заметны при малом объеме онтологии (тысячи объектов). Однако они выходят на передний план при использовании больших баз данных (например, Википедии или *Freebase*).

Рост онтологической омонимии

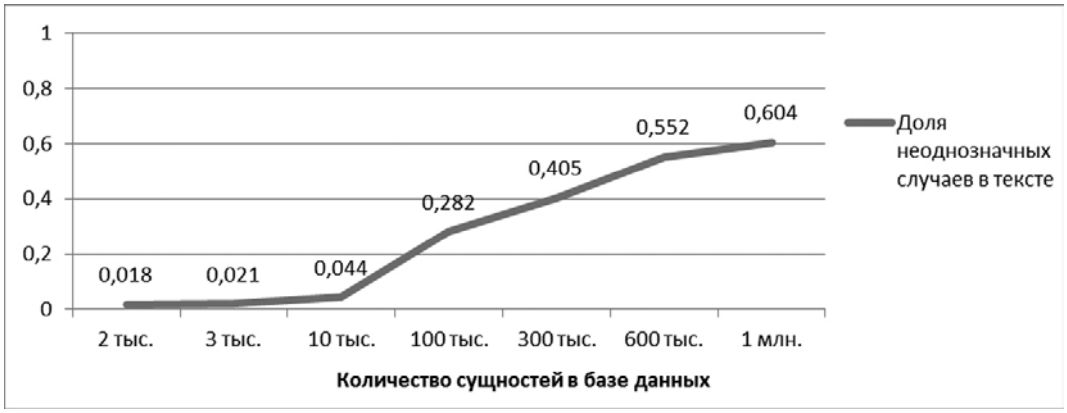

Рис. 1. *Возрастание степени неоднозначности имен в тексте при увеличении размера*

На Рис. 1 представлена диаграмма распределения количества сущностей БД, соответствующих одному и тому же имени в тексте, в зависимости от размера базы. Как видно, если в БД присутствует лишь 10 тыс. объектов, то количество случаев неоднозначности в тексте не превышает 5%. Однако уже при увеличении размера базы до 100 тыс. сущностей частота неоднозначных случаев приближается к 29%, а при миллионе сущностей – к 61%. Так, в онтологии *Freebase* имени *Washington* соответствуют 77 населенных пунктов.

Проблема языковой омонимии также значительно возрастает при использовании больших объемов данных. В онтологии *Freebase* некоторые географические объекты имеют имена *The*, *N/A*, *Bank*, *Friday*, *Monday*, *New*, *June*, *March*, *May*, *Liberty* и т.д. Это может являться как результатом ошибки при заполнении онтологии, так и случайным совпадением имени. График возрастания частоты проблемы языковой омонимии представлен на Рис. 2.

Рост языковой омонимии в зависимости от размера БД

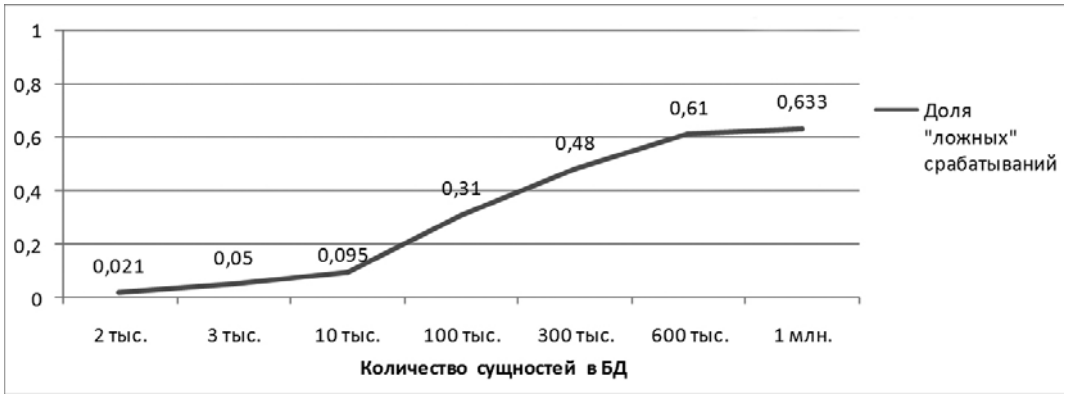

Рис. 2. *Рост частоты «ложных» имен в тексте (т.е. случаев, когда имя случайно совпадает с общеупотребительным словом) в зависимости от размера БД*

Предыдущие работы

Целью данной статьи является апробация метода, использующего априорную информацию о соответствии между именованной сущностью в тексте и географическим объектом в БД. Похожая попытка уже была предпринята в статьях А. Фейдера и др. [3] и Й. Хоффарта и др. [4]. Исследование А. Фейдера и др. базируется на ранге, приписываемого соответствию между сущностью и именем на основе результатов поиска по Википедии, однако, во-первых, оно проводилось на малом количестве объектов (500 соответствий между именем и онтологической сущностью); во-вторых, использовались тексты другого характера (веб-страницы из блогов, новостей и рассказов). В статье Й. Хоффарта и др. метод приписывает ранг на основе частоты соответствия между сущностью и именем, однако подробное описание отсутствует. Помимо этого, исследование Й. Хоффарта и др. также проводилось на малом количестве текстов (около 600). Таким образом, можно заключить, что метод априорной информации нуждается в дополнительном исследовании.

Метод априорной информации

Разработанный метод наиболее вероятного объекта выбирает для имени в тексте ту связь, которая чаще всего встречалась в обучающем корпусе. Т.е. если имени *Washington* в большинстве случаев соответствует одна и та же сущность, во всех остальных случаях будет выбираться именно эта сущность. Этот метод имеет два настраиваемых параметра:

1. Минимальная вероятность p для выбора сущности:

$$p = \frac{w}{N},$$

где w – количество случаев, когда сущность соответствовала данному имени, а N – количество упоминаний данного имени.

2. Минимальное количество упоминаний имени в тестовом корпусе (m).

Мы ожидаем, что оба параметра имеют прямую корреляцию с точностью и обратную корреляцию с полнотой. Помимо влияния указанных выше параметров рассматривалась связь размера обучающего корпуса с достижимым качеством.

Зависимость качества метода априорной информации от минимальной вероятности p

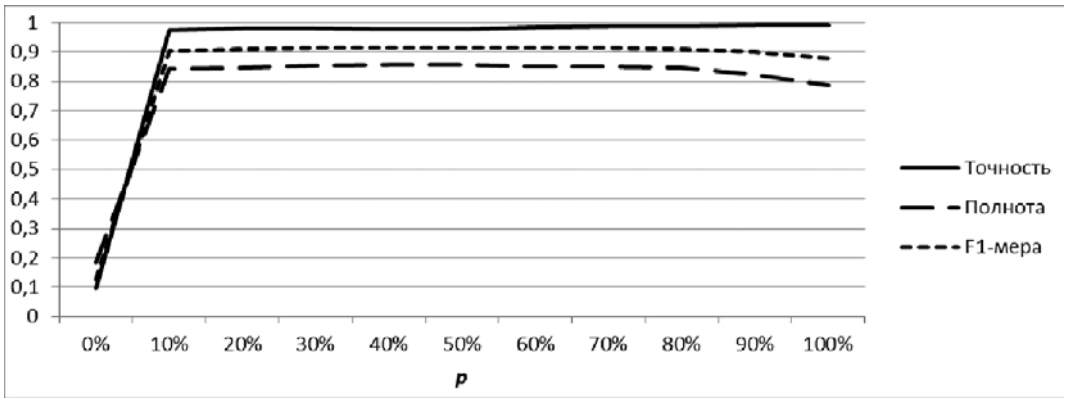

Рис. 3. *Зависимость качества метода априорной информации от минимальной вероятности p*

Как видно из Рис. 3, параметр p имеет одну вырожденную точку (ноль), когда качество работы метода оказывается значительно хуже других значений. Начиная от p=10% точность метода возрастает, достигая в пике 99,1% (что оправдывает наши ожидания), а полнота ведет себя несколько разнонаправленно, и лишь начиная с точки p=80% уверенно падает. Функция качества (F1-мера) достигает наибольшего значения в промежутке значения параметра p 30-50%.

Зависимость качества метода априорной информации от параметра *m*

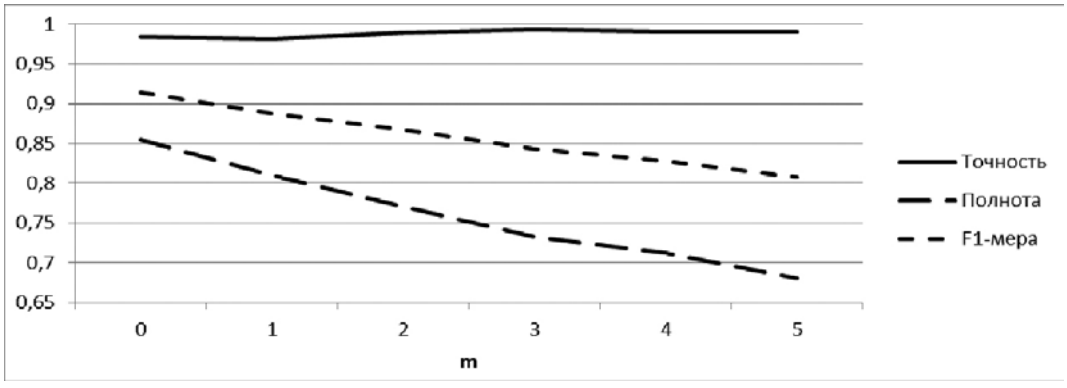

Рис. 4. *Зависимость качества метода априорной информации от параметра m*

Параметр *m* (минимальное количество упоминаний имени в тестовом корпусе) имеет ясный физический смысл: отсекая имена, встречающиеся редко, мы избавляемся от недостоверных случаев соответствия между именем и онтологической сущностью. Например, если в генеральном корпусе сущность E соответствует имени S в 40% случаев, имеется некоторый шанс, что в обучающий корпус попадет лишь малое количество употреблений имени, и статистика соответствия будет сильно искажена. Однако, как видно из Рис. 4, наше предположение оказалось скорее неверным. Увеличивая параметр *m*, мы отбрасываем не только недостоверные, но и малочастотные достоверные случаи, причем, судя по графику, последних значительно больше. На графике из Рис. 4 наблюдается и прямая корреляция с точностью (рост с 0,984 до 0,991), и обратная корреляция с полнотой (падение от 0,854 до 0,68). Качество работы метода априорной информации (F1-мера) падает при любых значениях параметра *m*, отличных от нуля. Теоретически, при очень больших размерах обучающего корпуса может образоваться число больше нуля, при котором F1-мера увеличится, однако на наших объемах в 32 тыс. гипотез ИС получить такое число не удалось.

Минимальный размер обучающего корпуса

Рассмотрим характер изменения качества метода в зависимости от размера обучающего корпуса. Единицей измерения размера корпуса разумно считать новостную статью, т.к. трудно представить себе рост корпуса с помощью другого механизма, кроме добавления новой статьи.

Зависимость качества метода априорной информации от размеров обучающего корпуса

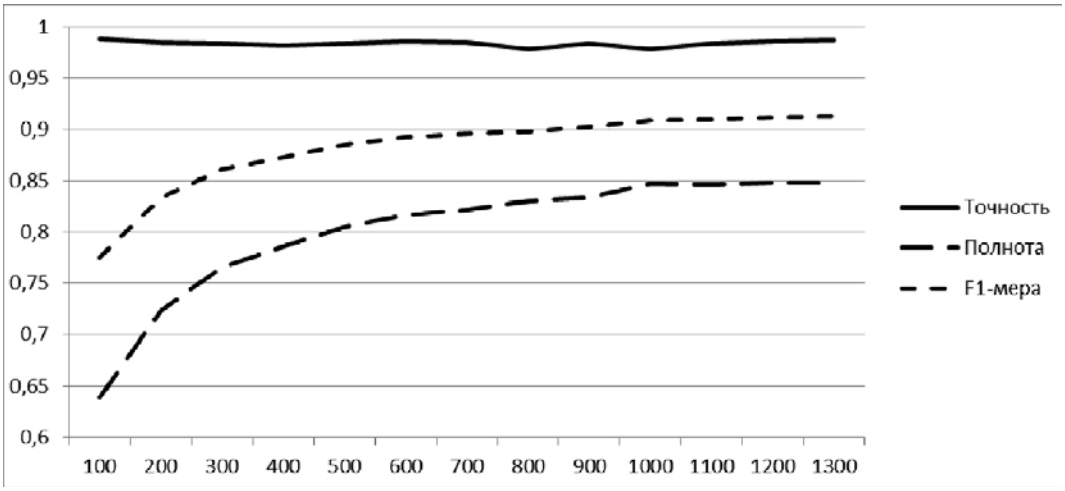

Рис. 5. *Зависимость качества метода априорной информации от размеров обучающего корпуса.*

На графике заметен логарифмический рост полноты при практически неизменной точности

Как видно из графика на Рис. 5, точность метода практически не меняется при изменении размеров обучающего корпуса. Небольшие колебания можно объяснить простой статистической погрешностью. Полнота же, напротив, показывает логарифмический рост в зависимости от количества статей в обучающем корпусе. Рост полноты стабилизируется на тысяче статей и в дальнейшем все больше затормаживается. Таким образом, можно утверждать, что для создания системы, использующей метод априорной информации, необходима, по меньшей мере, тысяча новостных статей.

Выводы

В статье был подробно рассмотрен метод разрешения языковой и онтологической омонимии именованных сущностей, основанный на априорной информации из обучающего корпуса. Данный метод имеет два настраиваемых параметра, с помощью которых можно регулировать итоговую полноту и точность: вероятность соответствия между сущностью и именем (p) и минимальное количество упоминаний имени в обучающем корпусе (m). Оба параметра имеют прямую корреляцию с точностью и обратную корреляцию с полнотой, наилучшие значения в нашем случае оказались 0 для параметра m и любое значение из интервала 30-50% для параметра p. Для качественного обучения при помощи системы, использующей метод априорной информации, необходимо минимум 1 тыс. статей. Система, использующая метод априорной информации, принимает верные решения только на сущностях, присутствовавших в обучающем корпусе. Расширение БД происходит за счет низкочастотных сущностей, и шанс, что те попадут в обучающий корпус, невелик. Поэтому использование метода априорной информации при большом размере БД приводит к низкому качеству работы системы.

Список литературы

1. **Антонов Е. С.** Как найти миллион // RSDN Magazine. СПб.: K-Press, 2011. № 1. С. 60-68.
2. **Cucerzan S.** Large Scale Named Entity Disambiguation Based on Wikipedia Data // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA: Association for Computational Linguistics, 2007. P. 708-716.

3. **Fader A., Soderland S., Etzioni O.** Scaling Wikipedia-Based Named Entity Disambiguation to Arbitrary Web Text // Proceedings of the WikiAI 09 - IJCAI Workshop: User Contributed Knowledge and Artificial Intelligence: an Evolving Synergy. Pasadena, CA: IJCAI Organization, 2009. P. 21-28.
4. **Hoffart J., Yosef M. A., Bordino I., Fürstenau H., Pinkal M., Spaniol M., Taneva B., Thater S., Weikum G.** Robust Disambiguation of Named Entities in Text // Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2011. P. 782–792.
5. **http://wiki.freebase.com/wiki/Main_Page**
6. **http://www.wikipedia.org**

A PRIORY INFORMATION AS WAY OF ONTOLOGICAL AND LANGUAGE HOMONYMY DISAMBIGUATION

Egor Sergeevich Antonov
Department of Theoretical and Applied Linguistics
Moscow State University named after M. V. Lomonosov
me@eantonov.name

The author considers the reasonability of a priory information use for the disambiguation of the language and ontological homonymy of named entities, by the material of marked corpus from 1700 English-language news articles verifies the strategy of choosing the most probable object with two adaptable parameters (the minimum probability of compliance, the minimum number of references in the corpus), and concludes that such a strategy allows achieving the high accuracy of homonymy disambiguation, but its use does not make sense for a large number of ontology objects because of low completeness.

Key words and phrases: named entities recognition; disambiguation of named entities homonymy; ontology; a priory information; geographic objects; news texts.

УДК 81’42
Филологические науки

Статья посвящена изучению феномена дискурса субботнего обращения президента США как одного из жанров ритуальной политической коммуникации. Основное внимание автор акцентирует на стратагемно-тактических особенностях данного вида дискурса. Установлено, что базовой стратегией дискурса субботнего обращения является стратегия интеграции как способ солидаризации, которая вербализуется через тактики диалогичности, комплиментарности и интертекстуальности.

Ключевые слова и фразы: политический дискурс; ритуальный жанр; дискурс обращения; стратегия интеграции; вербализация; тактика.

Ольга Вячеславовна Атьман, к. филол. н.
Кафедра профессиональной иноязычной коммуникации
Волгоградский государственный университет
olga-atman@yandex.ru

ВЕРБАЛИЗАЦИЯ СТРАТЕГИИ ИНТЕГРАЦИИ В ДИСКУРСЕ СУББОТНЕГО ОБРАЩЕНИЯ ПРЕЗИДЕНТА США©

Дискурс обращения президента не имеет аналога в российской политической культуре, напротив, в Соединённых Штатах он является одним из наиболее частотных жанров политического дискурса: за время своего четырёхлетнего пребывания на посту главы государства президент еженедельно выступает с субботней речью перед американской аудиторией. Например, с момента инаугурации 44-го президента США Барака Обамы, состоявшейся 20 января 2009 года, им произнесено более 180 субботних обращений к нации.

Политический дискурс уже давно рассматривается как сложный конгломерат интегральных/ритуальных, ориентационных и агональных жанров, типизируемых по характеру ведущей интенции. Исследуемый нами дискурс субботнего обращения президента США к нации является одним из интегральных жанров политического дискурса, поскольку жанры такого рода (к числу интегральных жанров мы также относим дискурс инаугурационного обращения, дискурс прощальной речи президента, дискурс рождественского обращения президента к нации) функционально подчинены интенции интеграции, т.е. намерению сплотить и объединить нацию.

Доминирующей стратегией дискурса субботнего обращения к нации мы считаем стратегию ***интеграции*** как способ кооперации и солидаризации с американскими слушателями – потенциальными избирателями президента, олицетворяющего собой государственную власть, как стремление к обеспечению верховной властью возможности мирного сосуществования различных социальных и этнических групп населения в единой системе национального государства. Именно этой стратегии подчинена вся коммуникативная организация ритуального события. Т. Н. Астафурова и А. В. Олянич справедливо отмечают, что «оратор как активный творец речи, выражающий свои эмоции, побуждающий слушателей к действиям, выполняет одну из