Лингвистические аспекты новых информационных технологий

Вычислительная морфология.

Алексей Андреевич Сорокин

ОТИПЛ МГУ,
осенний семестр 2017–2018 учебного года

Постановка задачи

Дан текст на естественном языке, требуется автоматически произвести морфоологический анализ входящих в него слов.

Его	DET
решение	NOUN, gender=Neut, case=Nom, number=Plur
задачи	NOUN, gender=Fem, case=Gen, number=Sing
было	AUX, tense=Past, aspect=Imp, number=Sing, gender=Neut
неправильным	ADJ, number=Sing, gender=Neut, case=Ins

Постановка задачи

Часто нужно дополнительно восстановить начальные формы (леммы) входящих в текст словоформ.

Его	DET	его
решение	NOUN, gender=Neut, case=Nom, number=Plur	решение
задачи	NOUN, gender=Fem, case=Gen, number=Sing	задача
было	AUX, tense=Past, aspect=Imp, number=Sing, gender=Neut	быть
неправильным	ADJ, number=Sing, gender=Neut, case=Ins	неправильный

Применение

- Переход к более глубинному представлению текста.
- Извлечение информации из текста (нужны лексемы, а не словоформы).

Применение

- Переход к более глубинному представлению текста.
- Извлечение информации из текста (нужны лексемы, а не словоформы).
- Подготовка текста для дальнейшего анализа (извлечение сущностей $\rightarrow$ нужны именные группы $\rightarrow$ нужно знать морфологические метки).

Применение

- Переход к более глубинному представлению текста.
- Извлечение информации из текста (нужны лексемы, а не словоформы).
- Подготовка текста для дальнейшего анализа (извлечение сущностей $\rightarrow$ нужны именные группы $\rightarrow$ нужно знать морфологические метки).
- Уточнение вероятностной модели текста, снижение разреженности (лексем меньше, чем словоформ).

Применение

- Переход к более глубинному представлению текста.
- Извлечение информации из текста (нужны лексемы, а не словоформы).
- Подготовка текста для дальнейшего анализа (извлечение сущностей $\rightarrow$ нужны именные группы $\rightarrow$ нужно знать морфологические метки).
- Уточнение вероятностной модели текста, снижение разреженности (лексем меньше, чем словоформ).
- Автоматическое пополнение, создание и верификация лексических ресурсов (корпуса, словари и т. д.).

Задачи вычислительной морфологии

- Морфологический анализ (на уровне предложений).

Задачи вычислительной морфологии

- Морфологический анализ (на уровне предложений).
- Лемматизация (на уровне предложений/отдельных слов).

Задачи вычислительной морфологии

- Морфологический анализ (на уровне предложений).
- Лемматизация (на уровне предложений/отдельных слов).
- Морфологический синтез.
- Автоматическое построение и пополнение парадигм.
- Автоматическое деление на морфемы.

Сложности в задачах вычислительной морфологии

- Неизбежное наличие несловарных форм.

Сложности в задачах вычислительной морфологии

- Неизбежное наличие несловарных форм.
- Даже для словарных форм – регулярная омонимия.

Сложности в задачах вычислительной морфологии

- Неизбежное наличие несловарных форм.
- Даже для словарных форм – регулярная омонимия.
- Большой объём данных (сотни тысяч – миллионы возможных словоформ).

Сложности в задачах вычислительной морфологии

- Неизбежное наличие несловарных форм.
- Даже для словарных форм – регулярная омонимия.
- Большой объём данных (сотни тысяч – миллионы возможных словоформ).
- Нерегулярные модели словоизменения.

Сложности в задачах вычислительной морфологии

- Неизбежное наличие несловарных форм.
- Даже для словарных форм – регулярная омонимия.
- Большой объём данных (сотни тысяч – миллионы возможных словоформ).
- Нерегулярные модели словоизменения.
- Недостаточный объём или качество существующих ресурсов.

Сложности в задачах вычислительной морфологии

- Неизбежное наличие несловарных форм.
- Даже для словарных форм – регулярная омонимия.
- Большой объём данных (сотни тысяч – миллионы возможных словоформ).
- Нерегулярные модели словоизменения.
- Недостаточный объём или качество существующих ресурсов.
- Несогласованность между ресурсами.

Возможные способы описания

- Для английского категории можно задать списком: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`:

12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural

Возможные способы описания

- Для английского категории можно задать списком: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`:

12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural

- Для (поли-)синтетических языков не годится: слишком много категорий.

Возможные способы описания

- Для английского категории можно задать списком: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`:

12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural

- Для (поли-)синтетических языков не годится: слишком много категорий.
- К тому же нужно отразить внутренюю структуру грамматического значения.

Возможные способы описания

- Для английского категории можно задать списком: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`:

12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural

- Для (поли-)синтетических языков не годится: слишком много категорий.
- К тому же нужно отразить внутренюю структуру грамматического значения.
- Основные подходы:
  - Позиционный (проект Multext-East для славянских языков).
  - Признаковый (Universal dependencies).

Позиционный подход

- Используется в проекте Multext-East для восточноевропейских языков (`http://nl.ijs.si/ME/`).
- Метка – последовательность букв.

Позиционный подход

- Используется в проекте Multext-East для восточноевропейских языков (`http://nl.ijs.si/ME/`).
- Метка – последовательность букв.
- Первая большая буква — часть речи (N — существ., V — глагол, и т. д.).
- Для большинства славянских языков 13 базовых частей речи.

Позиционный подход

- Используется в проекте Multext-East для восточноевропейских языков (`http://nl.ijs.si/ME/`).
- Метка – последовательность букв.
- Первая большая буква — часть речи (N — существ., V — глагол, и т. д.).
- Для большинства славянских языков 13 базовых частей речи.
- Остальные буквы отражают признаки:

человек	Ncmsny	**c**ommon noun, **m**asculine, **s**ingular, **n**ominative, animate (**y**es).
выиграла	Vmis-sfa-e-	**m**ain verb, **i**ndicative, past(**s**), **s**ingular, **f**eminine, **a**ctive voice, perfect (**e**)

Позиционный подход

- Используется в проекте Multext-East для восточноевропейских языков (`http://nl.ijs.si/ME/`).
- Метка – последовательность букв.
- Первая большая буква — часть речи (N — существ., V — глагол, и т. д.).
- Для большинства славянских языков 13 базовых частей речи.
- Остальные буквы отражают признаки:

человек	Ncmsny	**c**ommon noun, **m**asculine, **s**ingular, **n**ominative, animate (**y**es).
выиграла	Vmis-sfa-e-	**m**ain verb, **i**ndicative, past(**s**), **s**ingular, **f**eminine, **a**ctive voice, perfect (**e**)

- Недостаток: метки сильно зависят от языка и стандарта.

Признаковые описания

- Основной пример: проект Universal Dependencies `http://universaldependencies.org`.

Признаковые описания

- Основной пример: проект Universal Dependencies `http://universaldependencies.org`.
- Цель: создать универсальный формат синтаксической и морфологической разметки, допускающий расширение на новые языки и корпуса.

Признаковые описания

- Основной пример: проект Universal Dependencies `http://universaldependencies.org`.
- Цель: создать универсальный формат синтаксической и морфологической разметки, допускающий расширение на новые языки и корпуса.
- В основном автоматическая разметка с ручной верификацией.

Признаковые описания

- Основной пример: проект Universal Dependencies `http://universaldependencies.org`.
- Цель: создать универсальный формат синтаксической и морфологической разметки, допускающий расширение на новые языки и корпуса.
- В основном автоматическая разметка с ручной верификацией.
- На март 2017 года — 50 языков (версия 2.0).

Признаковые описания

- Основной пример: проект Universal Dependencies http://universaldependencies.org.
- Цель: создать универсальный формат синтаксической и морфологической разметки, допускающий расширение на новые языки и корпуса.
- В основном автоматическая разметка с ручной верификацией.
- На март 2017 года — 50 языков (версия 2.0).
- Метки задаются в формате CONLL-U http://universaldependencies.org/format.html.

Признаковые описания

- Основной пример: проект Universal Dependencies `http://universaldependencies.org`.
- Цель: создать универсальный формат синтаксической и морфологической разметки, допускающий расширение на новые языки и корпуса.
- В основном автоматическая разметка с ручной верификацией.
- На март 2017 года — 50 языков (версия 2.0).
- Метки задаются в формате CONLL-U `http://universaldependencies.org/format.html`.
- Две части у каждой метки: универсальная часть речи (UPOSTAG) и признаковое описание (FEATS).

Признаковые описания

- Основной пример: проект Universal Dependencies `http://universaldependencies.org`.
- Цель: создать универсальный формат синтаксической и морфологической разметки, допускающий расширение на новые языки и корпуса.
- В основном автоматическая разметка с ручной верификацией.
- На март 2017 года — 50 языков (версия 2.0).
- Метки задаются в формате CONLL-U `http://universaldependencies.org/format.html`.
- Две части у каждой метки: универсальная часть речи (UPOSTAG) и признаковое описание (FEATS).
- 17 универсальных частей речи:

ADJ	adjective	INTJ	interjection	PUNCT	punctuation
ADP	adposition	NOUN	noun	SCONJ	subordinating conjunction
ADV	adverb	NUM	numeral	SYM	symbol
AUX	auxiliary	PART	particle	VERB	verb
CCONJ	coordinating conjunction	PRON	pronoun	X	other
DET	determiner	PROPN	proper noun		

Признаковые описания

- 21 признак: 6 лексических и 15 морфологических.

Признаковые описания

- 21 признак: 6 лексических и 15 морфологических.
- Лексические признаки: PronType, NumType, Poss, Reflex, Foreign, Abbr.

Признаковые описания

- 21 признак: 6 лексических и 15 морфологических.
- Лексические признаки: PronType, NumType, Poss, Reflex, Foreign, Abbr.
- Морфологические признаки:
  - Именные: Gender, Animacy, Number, Case, Degree, Definite.

Признаковые описания

- 21 признак: 6 лексических и 15 морфологических.
- Лексические признаки: PronType, NumType, Poss, Reflex, Foreign, Abbr.
- Морфологические признаки:
  - Именные: Gender, Animacy, Number, Case, Degree, Definite.
  - Глагольные: VerbForm, Mood, Tense, Aspect, Voice, Evident, Polarity, Person, Polite.

Признаковые описания

- 21 признак: 6 лексических и 15 морфологических.
- Лексические признаки: PronType, NumType, Poss, Reflex, Foreign, Abbr.
- Морфологические признаки:
  - Именные: Gender, Animacy, Number, Case, Degree, Definite.
  - Глагольные: VerbForm, Mood, Tense, Aspect, Voice, Evident, Polarity, Person, Polite.
- Если в языке нет категории, соответствующему признаку просто не присваивается значение

Признаковые описания

- 21 признак: 6 лексических и 15 морфологических.
- Лексические признаки: PronType, NumType, Poss, Reflex, Foreign, Abbr.
- Морфологические признаки:
  - Именные: Gender, Animacy, Number, Case, Degree, Definite.
  - Глагольные: VerbForm, Mood, Tense, Aspect, Voice, Evident, Polarity, Person, Polite.
- Если в языке нет категории, соответствующему признаку просто не присваивается значение
- Значения категорий зависят от языка, но стандартизированы.

Особенности русского языка

- Большое количество неэквивалентных форматов: Universal Dependencies, НКРЯ, OpenCorpora, ГИКРЯ (вариант MSD), MSD, MyStem (вариант НКРЯ).

Особенности русского языка

- Большое количество неэквивалентных форматов: Universal Dependencies, НКРЯ, OpenCorpora, ГИКРЯ (вариант MSD), MSD, MyStem (вариант НКРЯ).
- Признаковые описания более удобны.
- Пограничные категории:
  - Причастия: прилагательное или форма глагола.

Особенности русского языка

- Большое количество неэквивалентных форматов: Universal Dependencies, НКРЯ, OpenCorpora, ГИКРЯ (вариант MSD), MSD, MyStem (вариант НКРЯ).
- Признаковые описания более удобны.
- Пограничные категории:
  - Причастия: прилагательное или форма глагола.
  - Предикативы vs наречия vs краткие прилагательные — где граница.

Особенности русского языка

- Большое количество неэквивалентных форматов: Universal Dependencies, НКРЯ, OpenCorpora, ГИКРЯ (вариант MSD), MSD, MyStem (вариант НКРЯ).
- Признаковые описания более удобны.
- Пограничные категории:
  - Причастия: прилагательное или форма глагола.
  - Предикативы vs наречия vs краткие прилагательные — где граница.
  - Вид — словоизменительная или классифицирующая категория.

Особенности русского языка

- Большое количество неэквивалентных форматов: Universal Dependencies, НКРЯ, OpenCorpora, ГИКРЯ (вариант MSD), MSD, MyStem (вариант НКРЯ).
- Признаковые описания более удобны.
- Пограничные категории:
  - Причастия: прилагательное или форма глагола.
  - Предикативы vs наречия vs краткие прилагательные — где граница.
  - Вид — словоизменительная или классифицирующая категория.
  - Двувидовые глаголы.
  - Классификация местоимений.

Особенности русского языка

- Большое количество неэквивалентных форматов: Universal Dependencies, НКРЯ, OpenCorpora, ГИКРЯ (вариант MSD), MSD, MyStem (вариант НКРЯ).
- Признаковые описания более удобны.
- Пограничные категории:
  - Причастия: прилагательное или форма глагола.
  - Предикативы vs наречия vs краткие прилагательные — где граница.
  - Вид — словоизменительная или классифицирующая категория.
  - Двувидовые глаголы.
  - Классификация местоимений.
- Попытка унификации: MorphoRuEval-2017 (Sorokin et al., 2017).

Постановка задачи

- Дана последовательность слов (токенов) $\mathbf{w}_{1,n} = w_1 \dots w_n$.
- Требуется найти соответствующую последовательлность морфологических меток $\mathbf{t}_{1,n} = t_1 \dots t_n$.

Постановка задачи

- Дана последовательность слов (токенов) $\mathbf{w}_{1,n} = w_1 \dots w_n$.
- Требуется найти соответствующую последовательлность морфологических меток $\mathbf{t}_{1,n} = t_1 \dots t_n$.
- Ищется наиболее вероятная последовательность:

$$\mathbf{t} = t_1 \dots t_n = \text{argmax}_{\mathbf{t}}\, p(t_1 \dots t_n | w_1 \dots w_n)$$

Постановка задачи

- Дана последовательность слов (токенов) $\mathbf{w}_{1,n} = w_1 \dots w_n$.
- Требуется найти соответствующую последовательлность морфологических меток $\mathbf{t}_{1,n} = t_1 \dots t_n$.
- Ищется наиболее вероятная последовательность:

$$\mathbf{t} = t_1 \dots t_n = \text{argmax}_{\mathbf{t}}\, p(t_1 \dots t_n | w_1 \dots w_n)$$

- Метки выбираются из заранее известного множества $\mathcal{T}$ (обычно $|\mathcal{T}| \approx 100 - 1000$).

Постановка задачи

- Дана последовательность слов (токенов) $\mathbf{w}_{1,n} = w_1 \dots w_n$.
- Требуется найти соответствующую последовательльность морфологических меток $\mathbf{t}_{1,n} = t_1 \dots t_n$.
- Ищется наиболее вероятная последовательность:

$$\mathbf{t} = t_1 \dots t_n = \text{argmax}_{\mathbf{t}}\, p(t_1 \dots t_n | w_1 \dots w_n)$$

- Метки выбираются из заранее известного множества $\mathcal{T}$ (обычно $|\mathcal{T}| \approx 100 - 1000$).
- Вспомогательные ресурсы:
  - Морфологические словари.
  - Корпуса с морфологической разметкой.

Модель “канала связи”

- Формально,

$$\mathbf{t} = t_1 \dots t_n = \text{argmax}_{\mathbf{t}}\, p(t_1 \dots t_n | w_1 \dots w_n)$$

Модель “канала связи”

- Формально,

$$\mathbf{t} = t_1 \dots t_n = \text{argmax}_{\mathbf{t}}\, p(t_1 \dots t_n | w_1 \dots w_n)$$

- Преобразуем по формуле Байеса:

$$\mathbf{t} = \text{argmax}_{\mathbf{t}}\, p(\mathbf{t}|\mathbf{w}) = \text{argmax}_{\mathbf{t}}\, \frac{p(\mathbf{w}|\mathbf{t})p(\mathbf{t})}{p(\mathbf{w})}$$

Модель “канала связи”

- Формально,

$$\mathbf{t} = t_1 \dots t_n = \operatorname{argmax}_{\mathbf{t}} p(t_1 \dots t_n | w_1 \dots w_n)$$

- Преобразуем по формуле Байеса:

$$\mathbf{t} = \operatorname{argmax}_{\mathbf{t}} p(\mathbf{t}|\mathbf{w}) = \operatorname{argmax}_{\mathbf{t}} \frac{p(\mathbf{w}|\mathbf{t})p(\mathbf{t})}{p(\mathbf{w})}$$

- $\mathbf{w}$ заранее дана — точка максимума не зависит от $p(\mathbf{w})$.

$$\mathbf{t} = p(\mathbf{w}|\mathbf{t})p(\mathbf{t})$$

Модель “канала связи”

- Формально,

$$\mathbf{t} = t_1 \dots t_n = \operatorname{argmax}_{\mathbf{t}} p(t_1 \dots t_n | w_1 \dots w_n)$$

- Преобразуем по формуле Байеса:

$$\mathbf{t} = \operatorname{argmax}_{\mathbf{t}} p(\mathbf{t}|\mathbf{w}) = \operatorname{argmax}_{\mathbf{t}} \frac{p(\mathbf{w}|\mathbf{t})p(\mathbf{t})}{p(\mathbf{w})}$$

- $\mathbf{w}$ заранее дана — точка максимума не зависит от $p(\mathbf{w})$.

$$\mathbf{t} = p(\mathbf{w}|\mathbf{t})p(\mathbf{t})$$

- $p(\mathbf{t})$ — вероятность последовательности меток (можно считать по энграммной модели).

Модель “канала связи”

- Формально,

$$\mathbf{t} = t_1 \dots t_n = \operatorname{argmax}_{\mathbf{t}} p(t_1 \dots t_n | w_1 \dots w_n)$$

- Преобразуем по формуле Байеса:

$$\mathbf{t} = \operatorname{argmax}_{\mathbf{t}} p(\mathbf{t}|\mathbf{w}) = \operatorname{argmax}_{\mathbf{t}} \frac{p(\mathbf{w}|\mathbf{t})p(\mathbf{t})}{p(\mathbf{w})}$$

- $\mathbf{w}$ заранее дана — точка максимума не зависит от $p(\mathbf{w})$.

$$\mathbf{t} = p(\mathbf{w}|\mathbf{t})p(\mathbf{t})$$

- $p(\mathbf{t})$ — вероятность последовательности меток (можно считать по энграммной модели).
- $p(\mathbf{w}|\mathbf{t})$ — вероятность “породить” данную последовательность слов $\mathbf{w}$.

Модель канала связи: объяснение

- $p(\mathbf{t})$ отвечает за грамматику (с помощью контекста):

Модель канала связи: объяснение

- $p(\mathbf{t})$ отвечает за грамматику (с помощью контекста):

  он решил пример за 5 *минут*
- Вероятность последовательности $\text{PRON VERB NOUN,case=Acc}$ выше, чем альтернативных вариантов.

Модель канала связи: объяснение

- $p(\mathbf{t})$ отвечает за грамматику (с помощью контекста):

  он решил пример за 5 *минут*
- Вероятность последовательности $\text{PRON VERB NOUN,case=Acc}$ выше, чем альтернативных вариантов.
- Непосредственный контекст не всегда помогает:

  Помещение было набито битком

Модель канала связи: объяснение

- $p(\mathbf{t})$ отвечает за грамматику (с помощью контекста):

  он решил пример за 5 *минут*
- Вероятность последовательности $\text{PRON VERB NOUN,case=Acc}$ выше, чем альтернативных вариантов.
- Непосредственный контекст не всегда помогает:

  Помещение было набито битком
- Однако *битком* — скорее всего наречие.

Модель канала связи: объяснение

- $p(\mathbf{t})$ отвечает за грамматику (с помощью контекста):

  он решил пример за 5 *минут*
- Вероятность последовательности $\text{PRON VERB NOUN,case=Acc}$ выше, чем альтернативных вариантов.
- Непосредственный контекст не всегда помогает:

  Помещение было набито битком
- Однако *битком* — скорее всего наречие.
- Формально, $p(\text{ADV}|\textit{битком}) \gg p(\text{NOUN}|\textit{битком})$.

Модель канала связи: объяснение

- $p(\mathbf{t})$ отвечает за грамматику (с помощью контекста):

  он решил пример за 5 *минут*
- Вероятность последовательности $\text{PRON VERB NOUN,case=Acc}$ выше, чем альтернативных вариантов.
- Непосредственный контекст не всегда помогает:

  Помещение было набито битком
- Однако *битком* — скорее всего наречие.
- Формально, $p(\text{ADV}|\textit{битком}) \gg p(\text{NOUN}|\textit{битком})$.
- Как следствие, $p(\textit{битком}|\text{ADV}) > p(\textit{битком}|\text{NOUN})$.

Марковская модель: постановка задачи

- Нам нужно оценить $p(\mathbf{w}|\mathbf{t})$.
- Предположение о независимости: w_i не зависит от других меток, кроме t_i.

Марковская модель: постановка задачи

- Нам нужно оценить $p(\mathbf{w}|\mathbf{t})$.
- Предположение о независимости: w_i не зависит от других меток, кроме t_i.
- Получаем:

$$p(\mathbf{w}|\mathbf{t}) = p(w_1 \dots w_n|t_1 \dots t_n) = p(w_1|t_1) \dots p(w_n|t_n)$$

Марковская модель: постановка задачи

- Нам нужно оценить $p(\mathbf{w}|\mathbf{t})$.
- Предположение о независимости: w_i не зависит от других меток, кроме t_i.
- Получаем:

$$p(\mathbf{w}|\mathbf{t}) = p(w_1 \dots w_n|t_1 \dots t_n) = p(w_1|t_1) \dots p(w_n|t_n)$$

- Теперь нужно оценить $p(w|t)$ (*лексическая вероятность*).
- Можно посчитать долю слова w среди слов с меткой t.

Марковская модель: постановка задачи

- Нам нужно оценить $p(\mathbf{w}|\mathbf{t})$.
- Предположение о независимости: w_i не зависит от других меток, кроме t_i.
- Получаем:

$$p(\mathbf{w}|\mathbf{t}) = p(w_1 \dots w_n|t_1 \dots t_n) = p(w_1|t_1) \dots p(w_n|t_n)$$

- Теперь нужно оценить $p(w|t)$ (*лексическая вероятность*).
- Можно посчитать долю слова w среди слов с меткой t.
- Не годится для слов, не встречавшихся в корпусе.

Марковская модель: постановка задачи

- Нам нужно оценить $p(\mathbf{w}|\mathbf{t})$.
- Предположение о независимости: w_i не зависит от других меток, кроме t_i.
- Получаем:

$$p(\mathbf{w}|\mathbf{t}) = p(w_1 \dots w_n|t_1 \dots t_n) = p(w_1|t_1) \dots p(w_n|t_n)$$

- Теперь нужно оценить $p(w|t)$ (*лексическая вероятность*).
- Можно посчитать долю слова w среди слов с меткой t.
- Не годится для слов, не встречавшихся в корпусе.
- Для слов из корпуса не учитывает информацию из словаря.

Лексические вероятности

- Вместо $p(w|t)$ удобнее оценивать $p(t|w)$.

Лексические вероятности

- Вместо $p(w|t)$ удобнее оценивать $p(t|w)$.
- Для словарных и корпусных слов:

$$p(w|t) = \frac{p(t|w)p(w)}{p(t)} \sim \frac{p(t|w)}{p(t)}$$

Лексические вероятности

- Вместо $p(w|t)$ удобнее оценивать $p(t|w)$.
- Для словарных и корпусных слов:

$$p(w|t) = \frac{p(t|w)p(w)}{p(t)} \sim \frac{p(t|w)}{p(t)}$$

- $p(t)$ — априорная вероятность метки t (её доля в корпусе/словаре).
- $p(t|w)$ — вероятность метки t для данного слова w.

Лексические вероятности

- Вместо $p(w|t)$ удобнее оценивать $p(t|w)$.
- Для словарных и корпусных слов:

$$p(w|t) = \frac{p(t|w)p(w)}{p(t)} \sim \frac{p(t|w)}{p(t)}$$

- $p(t)$ — априорная вероятность метки t (её доля в корпусе/словаре).
- $p(t|w)$ — вероятность метки t для данного слова w.
- Можно положить $p(t|w) \approx \frac{c(t,w)}{c(w)}$. Для $w = \textit{битком}$:

t	$c(t,w)$	$c(t)$	$\frac{c(t,\textit{битком})}{c(\textit{битком})}$	$p(w\|t) \sim \frac{p(t\|w)}{p(t)}$
ADV	20	10000	$\frac{20}{21}$	$0.8 \sim \frac{20}{21 \cdot 10000}$
NOUN, $case = Acc, \ldots$	1	2000	$\frac{1}{21}$	$0.2 \sim \frac{1}{21 \cdot 2000}$

Лексические вероятности

- Вместо $p(w|t)$ удобнее оценивать $p(t|w)$.
- Для словарных и корпусных слов:

$$p(w|t) = \frac{p(t|w)p(w)}{p(t)} \sim \frac{p(t|w)}{p(t)}$$

- $p(t)$ — априорная вероятность метки t (её доля в корпусе/словаре).
- $p(t|w)$ — вероятность метки t для данного слова w.
- Можно положить $p(t|w) \approx \frac{c(t,w)}{c(w)}$. Для $w = \textit{битком}$:

t	$c(t,w)$	$c(t)$	$\frac{c(t,\textit{битком})}{c(\textit{битком})}$	$p(w\|t) \sim \frac{p(t\|w)}{p(t)}$
ADV	20	10000	$\frac{20}{21}$	$0.8 \sim \frac{20}{21 \cdot 10000}$
NOUN, $case = Acc, \ldots$	1	2000	$\frac{1}{21}$	$0.2 \sim \frac{1}{21 \cdot 2000}$

- Снова не учитывается словарная информация!

Лексические вероятности

- Будем считать, что все словарные метки слова w дополнительно входят в корпус α раз (например, $\alpha = 0.5$)

Лексические вероятности

- Будем считать, что все словарные метки слова w дополнительно входят в корпус α раз (например, $\alpha = 0.5$)
- Тогда получим:

$$p(t|w) = \frac{c(t,w) + \alpha}{c(w) + \alpha|\mathcal{T}(w)|}$$

- $\mathcal{T}(w)$ — множество словарных меток слова w.

Лексические вероятности

- Будем считать, что все словарные метки слова w дополнительно входят в корпус α раз (например, $\alpha = 0.5$)
- Тогда получим:

$$p(t|w) = \frac{c(t,w) + \alpha}{c(w) + \alpha|\mathcal{T}(w)|}$$

- $\mathcal{T}(w)$ — множество словарных меток слова w.
- Для несловарных слов $p(t|w)$ считается непосредственно (машинное обучение).

Лексические вероятности

- Будем считать, что все словарные метки слова w дополнительно входят в корпус α раз (например, $\alpha = 0.5$)
- Тогда получим:

$$p(t|w) = \frac{c(t,w) + \alpha}{c(w) + \alpha|\mathcal{T}(w)|}$$

- $\mathcal{T}(w)$ — множество словарных меток слова w.
- Для несловарных слов $p(t|w)$ считается непосредственно (машинное обучение).
- Возможные признаки: суффиксы слова, символьные энграммы, капитализация.

Переборный алгоритм

Число вариантов быстро растёт с длиной предложения:

Его	DET
	PRON,gender=Masc,case=Gen
	PRON,gender=Masc,case=Acc
	PRON,gender=Neut,case=Gen
	PRON,gender=Neut,case=Acc
решение	NOUN,case=Nom
	NOUN,case=Acc
задачи	NOUN,number=Sing,case=Gen
	NOUN,number=Plur,case=Nom
	NOUN,number=Plur,case=Acc
было	AUX
	PART
неправильным	ADJ,number=Sing,case=Ins,gender=Masc
	ADJ,number=Sing,case=Ins,gender=Neut
	ADJ,number=Plur,case=Dat

Переборный алгоритм

- Для предложения длины 5 получается $5 * 2 * 3 * 2 * 3 = 180$ вариантов разбора.

Его	решение	задачи	было	неправильным
5	2	3	2	3

Переборный алгоритм

- Для предложения длины 5 получается $5 * 2 * 3 * 2 * 3 = 180$ вариантов разбора.

Его	решение	задачи	было	неправильным
5	2	3	2	3

- Рост числа вариантов экспоненциальный, т. е. для длины 10 возможно до ~ 10000 вариантов.

Переборный алгоритм

- Для предложения длины 5 получается $5 * 2 * 3 * 2 * 3 = 180$ вариантов разбора.

Его	решение	задачи	было	неправильным
5	2	3	2	3

- Рост числа вариантов экспоненциальный, т. е. для длины 10 возможно до ~ 10000 вариантов.
- Переборный алгоритм неприменим!

Марковская модель

- Пусть $p(\mathbf{t})$ задаётся триграммной моделью, т. е.

$$p(t_1 \dots t_n) = p(t_1)p(t_2|t_1)p(t_3|t_1 t_2) \dots p(t_n|t_{n-2} t_{n-1})$$

Марковская модель

- Пусть $p(\mathbf{t})$ задаётся триграммной моделью, т. е.

$$p(t_1 \dots t_n) = p(t_1)p(t_2|t_1)p(t_3|t_1 t_2) \dots p(t_n|t_{n-2} t_{n-1})$$

- Перепишем вероятность $p(\mathbf{t}|\mathbf{w}) \sim p(\mathbf{w}|\mathbf{t})p(\mathbf{t})$:

$$\begin{aligned} p(\mathbf{t}|\mathbf{w}) &\sim p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = \prod_{i=1}^{n} p(w_i|t_i) \prod_{i=1}^{n} p(t_i|t_{i-2} t_{i-1}) \\ &= \prod_{i=1}^{n} p(w_i|t_i)p(t_i|t_{i-2} t_{i-1}) \end{aligned}$$

Марковская модель

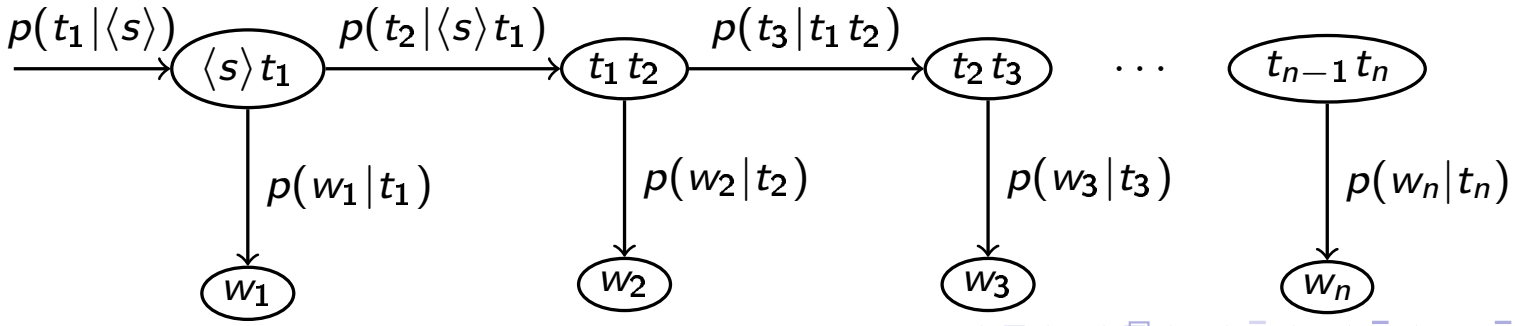
- Пусть $p(\mathbf{t})$ задаётся триграммной моделью, т. е.

$$p(t_1 \dots t_n) = p(t_1)p(t_2|t_1)p(t_3|t_1 t_2) \dots p(t_n|t_{n-2} t_{n-1})$$

- Перепишем вероятность $p(\mathbf{t}|\mathbf{w}) \sim p(\mathbf{w}|\mathbf{t})p(\mathbf{t})$:

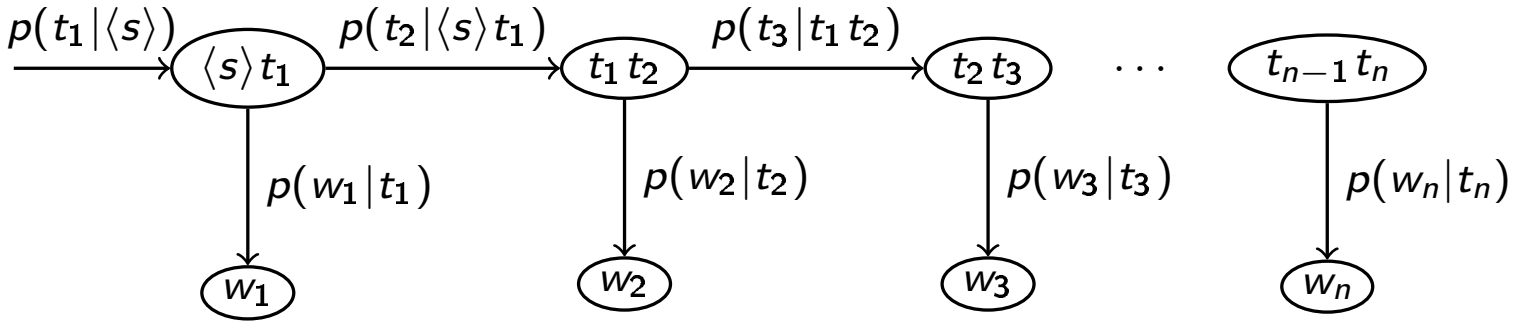
$$\begin{aligned} p(\mathbf{t}|\mathbf{w}) &\sim p(\mathbf{w}|\mathbf{t})p(\mathbf{t}) = \prod_{i=1}^{n} p(w_i|t_i) \prod_{i=1}^{n} p(t_i|t_{i-2} t_{i-1}) \\ &= \prod_{i=1}^{n} p(w_i|t_i)p(t_i|t_{i-2} t_{i-1}) \end{aligned}$$

- Графически ($\langle s \rangle$ — маркер начала строки)



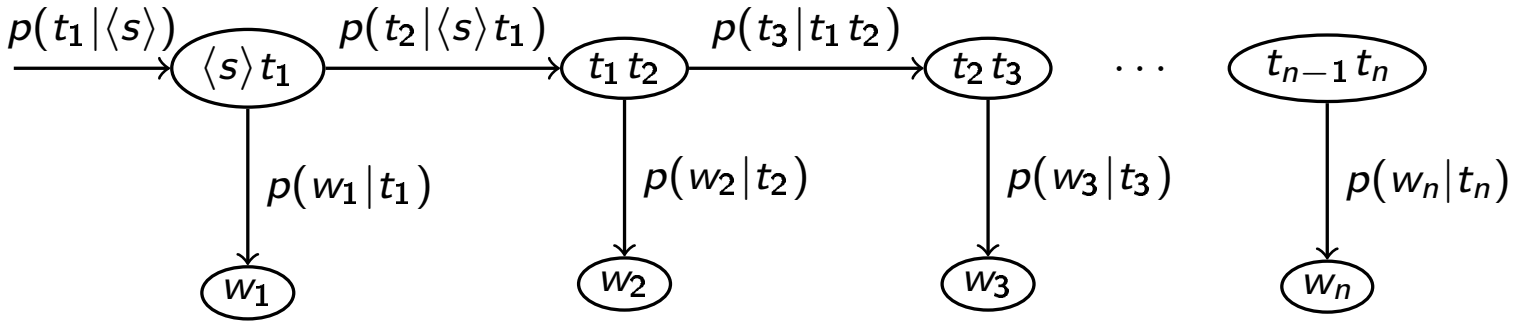
Марковская модель

- Графически ($\langle s \rangle$ — маркер начала строки)



Марковская модель

- Графически ($\langle s \rangle$ — маркер начала строки)

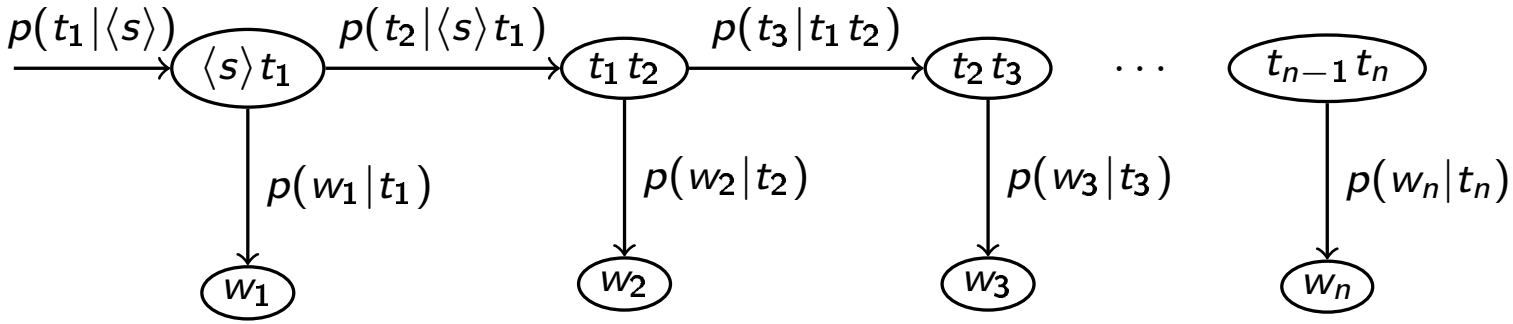


- Состояние модели в момент времени i: две последних метки

$$q(i) = t_{i-1} t_i.$$

Марковская модель

- Графически ($\langle s \rangle$ — маркер начала строки)



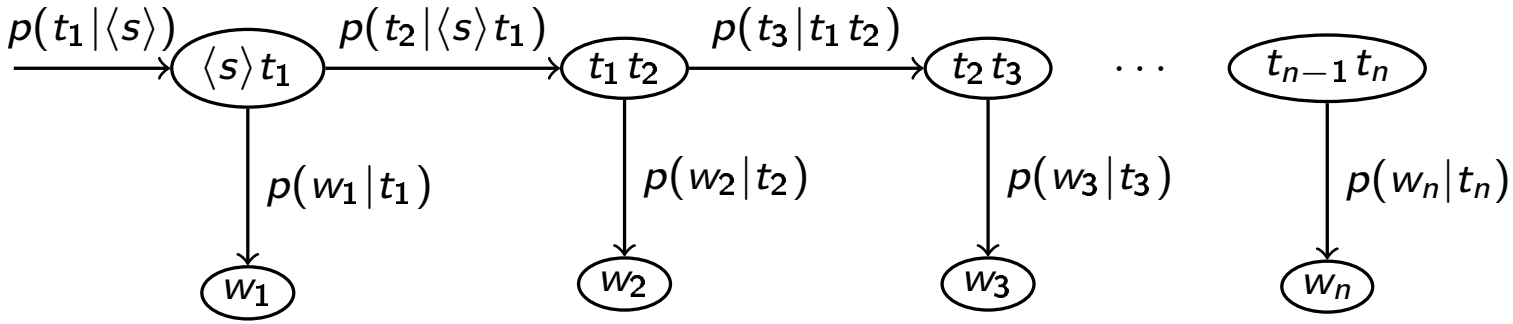
- Состояние модели в момент времени i: две последних метки

$$q(i) = t_{i-1}\,t_i.$$

- В каждый момент времени выдаётся слово w_i с вероятностью $p(w_i|q(i)) = p(w_i|t_i)$.

Марковская модель

- Графически ($\langle s \rangle$ — маркер начала строки)



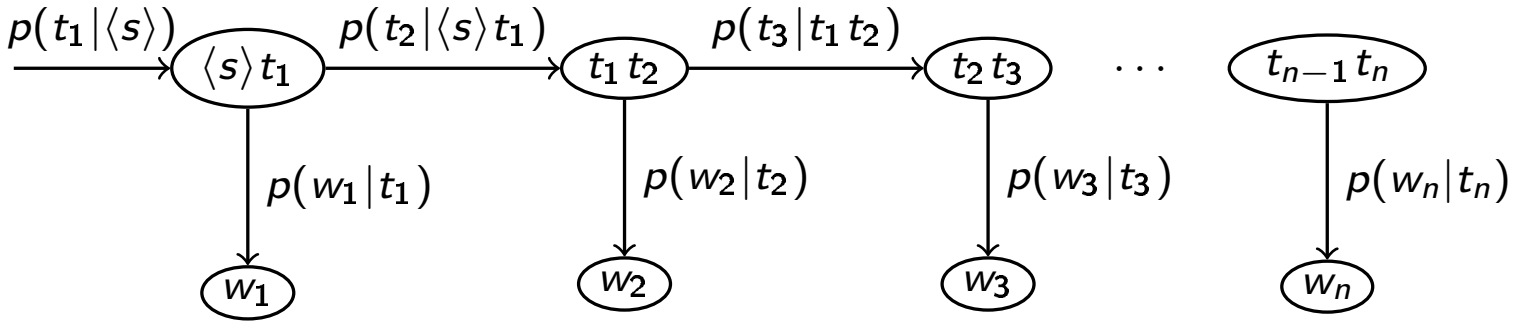
- Состояние модели в момент времени i: две последних метки

$$q(i) = t_{i-1}\, t_i.$$

- В каждый момент времени выдаётся слово w_i с вероятностью $p(w_i|q(i)) = p(w_i|t_i)$.
- Переход в следующее состояние $q(i+1) = t_i\, t_{i+1}$ происходит с вероятностью $p(t_{i+1}|t_{i-1}\, t_i)$.

Марковская модель

- Графически ($\langle s \rangle$ — маркер начала строки)



- Состояние модели в момент времени i: две последних метки

$$q(i) = t_{i-1} t_i.$$

- В каждый момент времени выдаётся слово w_i с вероятностью $p(w_i|q(i)) = p(w_i|t_i)$.
- Переход в следующее состояние $q(i+1) = t_i t_{i+1}$ происходит с вероятностью $p(t_{i+1}|t_{i-1} t_i)$.
- В начальный момент времени ничего не выдаётся (выдаётся фиктивный маркер $\langle s \rangle$).

Скрытые марковские модели

Определение

Скрытая марковская модель $\mathcal{M} = \langle Q, \Sigma, q_0, A, B \rangle$.

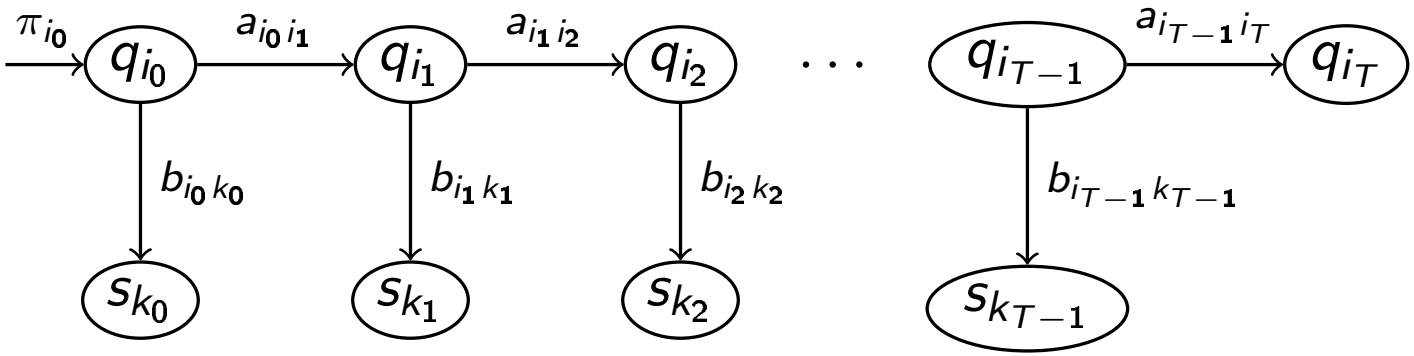
$$
\begin{array}{lcll}
Q & = & [q_1, \ldots, q_N] & — \text{ множество состояний,} \\
\Sigma & = & [s_1, \ldots, s_M] & — \text{ выходной алфавит,} \\
\pi & = & [\pi_1, \ldots, \pi_N] & — \text{ начальное распределение состояний,} \\
 & & A \in \mathbb{R}_{N \times N} & — \text{ вероятности перехода,} \\
 & & & a_{ij} = p(q(t) = q_j | q(t-1) = q_i) \\
 & & B \in \mathbb{R}_{N \times M} & — \text{ выходные вероятности,} \\
 & & & b_{ik} = p(X(t) = s_k | q(t) = q_i)
\end{array}
$$

$X(1), \ldots, X(t)$ — наблюдаемая последовательность,

$q(0), \ldots, q(t)$ — последовательность состояний.

Скрытые марковские модели

$$P(q_{i_0}, \ldots, q_{i_T}; s_{k_0}, \ldots, s_{k_{T-1}}) = \pi_{i_0} \prod_{t=1}^{T} a_{i_{t-1} i_t} b_{i_t k_t}$$



Нахождение наиболее вероятной последовательности состо

- Состояния марковской модели — пары морфологических меток (в триграммной модели).

Нахождение наиболее вероятной последовательности состо

- Состояния марковской модели — пары морфологических меток (в триграммной модели).
- Морфологический анализ сводится к нахождению наиболее вероятной последовательности состояний:

$$q_{i_0} \dots q_{i_n} = \text{argmax}_{q(0),\dots,q(n)}\, p(q(0), \dots, q(n)|X(1) \dots X(n))$$

Нахождение наиболее вероятной последовательности сост

- Состояния марковской модели — пары морфологических меток (в триграммной модели).
- Морфологический анализ сводится к нахождению наиболее вероятной последовательности состояний:

$$q_{i_0} \dots q_{i_n} = \text{argmax}_{q(0),\dots,q(n)}\, p(q(0), \dots, q(n)|X(1) \dots X(n))$$

- Основная идея: если зафиксировать предпоследнее состояние оптимального пути, то участок пути до него тоже должен быть оптимален.

Вероятность оптимальной последовательности состояний

- Запомним максимальную вероятность пути с $q(n) = i$:

$$A_{n,i} = \max_{q(0),\ldots,q(n-1),q(n)=i} p(q(0),\ldots,q(n)|X(1)\ldots X(n))$$

Вероятность оптимальной последовательности состояний

- Запомним максимальную вероятность пути с $q(n) = i$:

$$A_{n,i} = \max_{q(0),\ldots,q(n-1),q(n)=i} p(q(0),\ldots,q(n)|X(1)\ldots X(n))$$

- Тогда получаем:

$$A_{n,i} = \max_j A_{n-1,j} a_{ji} b_{i,k_n}$$

- k_n — номер n-го выходного символа (т. е. $X(n) = s_{k_n}$).

Вероятность оптимальной последовательности состояний

- Запомним максимальную вероятность пути с $q(n) = i$:

$$A_{n,i} = \max_{q(0),\ldots,q(n-1),q(n)=i} p(q(0),\ldots,q(n)|X(1)\ldots X(n))$$

- Тогда получаем:

$$A_{n,i} = \max_j A_{n-1,j} a_{ji} b_{i,k_n}$$

- k_n — номер n-го выходного символа (т. е. $X(n) = s_{k_n}$).
- В начальный момент времени $A_{0,i} = \pi_i$.
- Для морфологической марковской модели

$$\pi_i = 1, \text{ если } q_i = \langle s \rangle, \text{ иначе } p_i = 0.$$

Вероятность оптимальной последовательности состояний

- Запомним максимальную вероятность пути с $q(n) = i$:

$$A_{n,i} = \max_{q(0),\ldots,q(n-1),q(n)=i} p(q(0),\ldots,q(n)|X(1)\ldots X(n))$$

- Тогда получаем:

$$A_{n,i} = \max_j A_{n-1,j} a_{ji} b_{i,k_n}$$

- k_n — номер n-го выходного символа (т. е. $X(n) = s_{k_n}$).
- В начальный момент времени $A_{0,i} = \pi_i$.
- Для морфологической марковской модели

$$\pi_i = 1, \text{ если } q_i = \langle s \rangle, \text{ иначе } p_i = 0.$$

- Так за время $\sim n|Q|^2$ можно найти вероятность оптимальной последовательности состояний.

Вероятность оптимальной последовательности состояний

- Запомним максимальную вероятность пути с $q(n) = i$:

$$A_{n,i} = \max_{q(0),\ldots,q(n-1),q(n)=i} p(q(0),\ldots,q(n)|X(1)\ldots X(n))$$

- Тогда получаем:

$$A_{n,i} = \max_j A_{n-1,j} a_{ji} b_{i,k_n}$$

- k_n — номер n-го выходного символа (т. е. $X(n) = s_{k_n}$).
- В начальный момент времени $A_{0,i} = \pi_i$.
- Для морфологической марковской модели

$$\pi_i = 1, \text{ если } q_i = \langle s \rangle, \text{ иначе } p_i = 0.$$

- Так за время $\sim n|Q|^2$ можно найти вероятность оптимальной последовательности состояний.
- Но нам нужны сами состояния!

Нахождение оптимальной последовательности состояний

- При вычислении $A_{t,i}$ можно запомнить предыдущее оптимальное состояние:

$$\delta_{t,i} = \text{argmax}_j \, A_{t-1,j} a_{ji} b_{i,k_t}$$

Нахождение оптимальной последовательности состояний

- При вычислении $A_{t,i}$ можно запомнить предыдущее оптимальное состояние:

$$\delta_{t,i} = \text{argmax}_j \, A_{t-1,j} a_{ji} b_{i,k_t}$$

- Тогда оптимальные состояния восстанавливаются “обратным проходом” (backtracing):

$$\begin{aligned} q(n) &= \text{argmax}_i \, A_{n,i} \\ q(t-1) &= \delta_{t,q(t)} \end{aligned}$$

Нахождение оптимальной последовательности состояний

- При вычислении $A_{t,i}$ можно запомнить предыдущее оптимальное состояние:

$$\delta_{t,i} = \text{argmax}_j \, A_{t-1,j} a_{ji} b_{i,k_t}$$

- Тогда оптимальные состояния восстанавливаются “обратным проходом” (backtracing):

$$\begin{aligned} q(n) &= \text{argmax}_i \, A_{n,i} \\ q(t-1) &= \delta_{t,q(t)} \end{aligned}$$

- По последовательности состояний восстанавливаются оптимальные морфологические метки.

Недостатки марковских моделей

- Метки рассматриваются как единое целое, невозможно извлечь отдельные признаки.
- Сложнее выделить шаблоны согласования (вместо согласования по роду-числу-падежу — 3∗2∗6 отдельных согласований по каждому набору граммем).

Недостатки марковских моделей

- Метки рассматриваются как единое целое, невозможно извлечь отдельные признаки.
- Сложнее выделить шаблоны согласования (вместо согласования по роду-числу-падежу — 3∗2∗6 отдельных согласований по каждому набору граммем).
- Ограниченная память (чаще всего состояния — биграммы меток, триграмм уже слишком много).
- Следствие: не учитываются дистантные зависимости.

Недостатки марковских моделей

- Метки рассматриваются как единое целое, невозможно извлечь отдельные признаки.
- Сложнее выделить шаблоны согласования (вместо согласования по роду-числу-падежу — $3*2*6$ отдельных согласований по каждому набору граммем).
- Ограниченная память (чаще всего состояния — биграммы меток, триграмм уже слишком много).
- Следствие: не учитываются дистантные зависимости.
- t_n зависит только от w_n, t_{n-1}, t_{n-2}, но не от w_{n-1}.

Недостатки марковских моделей

- Метки рассматриваются как единое целое, невозможно извлечь отдельные признаки.
- Сложнее выделить шаблоны согласования (вместо согласования по роду-числу-падежу — 3∗2∗6 отдельных согласований по каждому набору граммем).
- Ограниченная память (чаще всего состояния — биграммы меток, триграмм уже слишком много).
- Следствие: не учитываются дистантные зависимости.
- t_n зависит только от w_n, t_{n-1}, t_{n-2}, но не от w_{n-1}.
- Однако лексемы влияют на морфологические показатели соседних:

обманул *друга* vs *соврал* *другу*
case=Gen case=Dat