ТЕОРЕТИКО-ИНФОРМАЦИОННЫЕ ГРАНИЦЫ И ЗАДАЧИ ЗАЩИТЫ ИНФОРМАЦИИ, СВЯЗАННЫЕ С ОБРАБОТКОЙ ЕСТЕСТВЕННОГО ЯЗЫКА

Мещеряков Р.В., д.т.н., проф., профессор РАН (ИПУ РАН).
Мельников С.Ю., д.ф.-м.н. (РУДН, ООО «Линфо»)

Работа поддержана грантом РНФ 24-11-00340 Исследование и разработка методов обработки слабоструктурированной информации на естественных языках в условиях сильных шумов для решения задач безопасности

16.05.24, г. Санкт-Петербург

ПЛАН ДОКЛАДА

1. Зашумление (искажение) текста
    1.1. О способах зашумления
    1.2. Задача восстановления зашумленного текста как оптимизационная задача
    1.3. Теоретико-информационные (шенноновские) границы уровня искажений, в которых возможно/невозможно восстановление зашумленного текста

2. Что можно понять из зашумленного текста, если его теоретически нельзя восстановить?
    2.1. Подход на основе локальных флуктуаций энтропии текста
    2.2. Подход на основе статистической оценки признаков текста, таких как язык, тематика, стиль, авторство и пр.

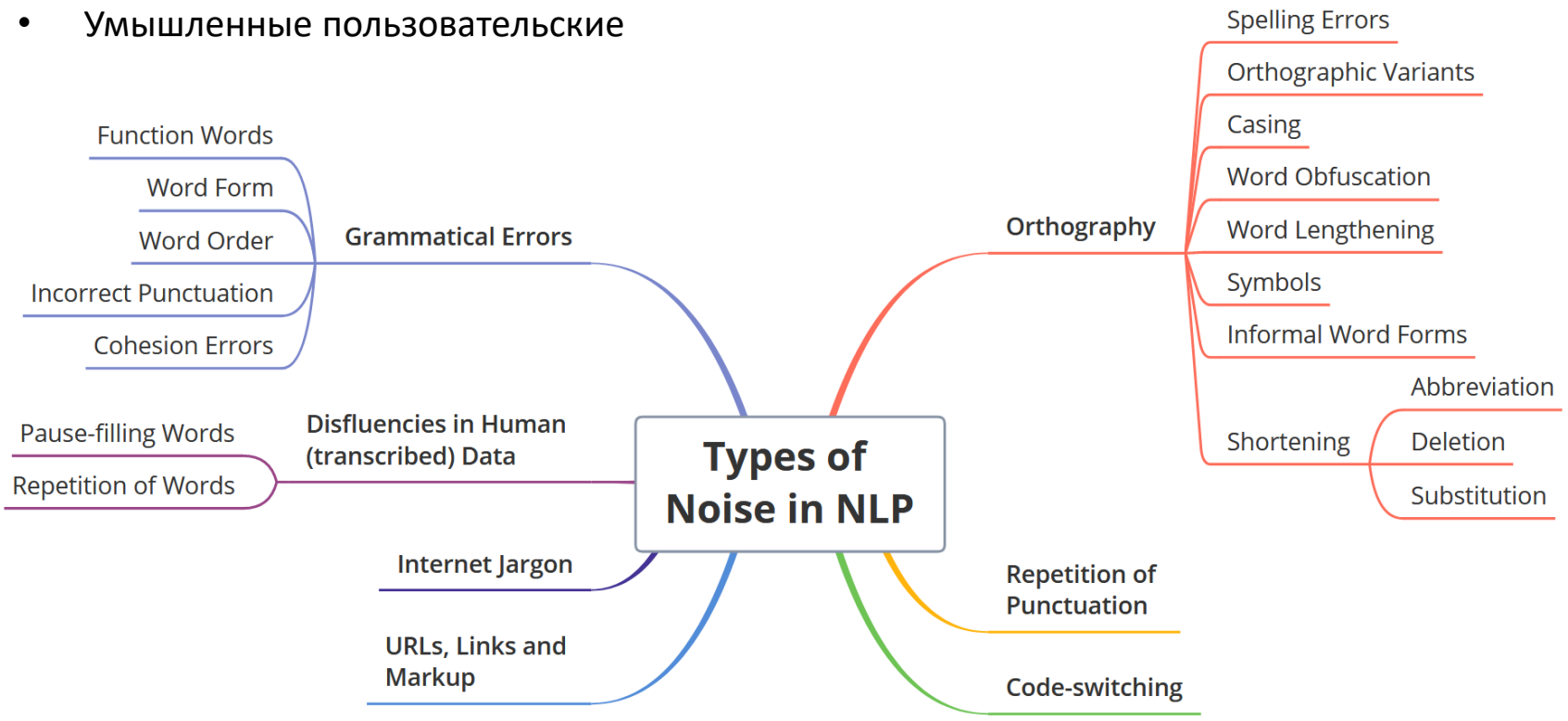
3. Опыты по пониманию речи в сильных шумах

4. Выводы

«ЕСТЕСТВЕННЫЕ» ИСКАЖЕНИЯ ТЕКСТА

Искажения в тексте:

- Случайные
- В результате распознавания
- Умышленные пользовательские

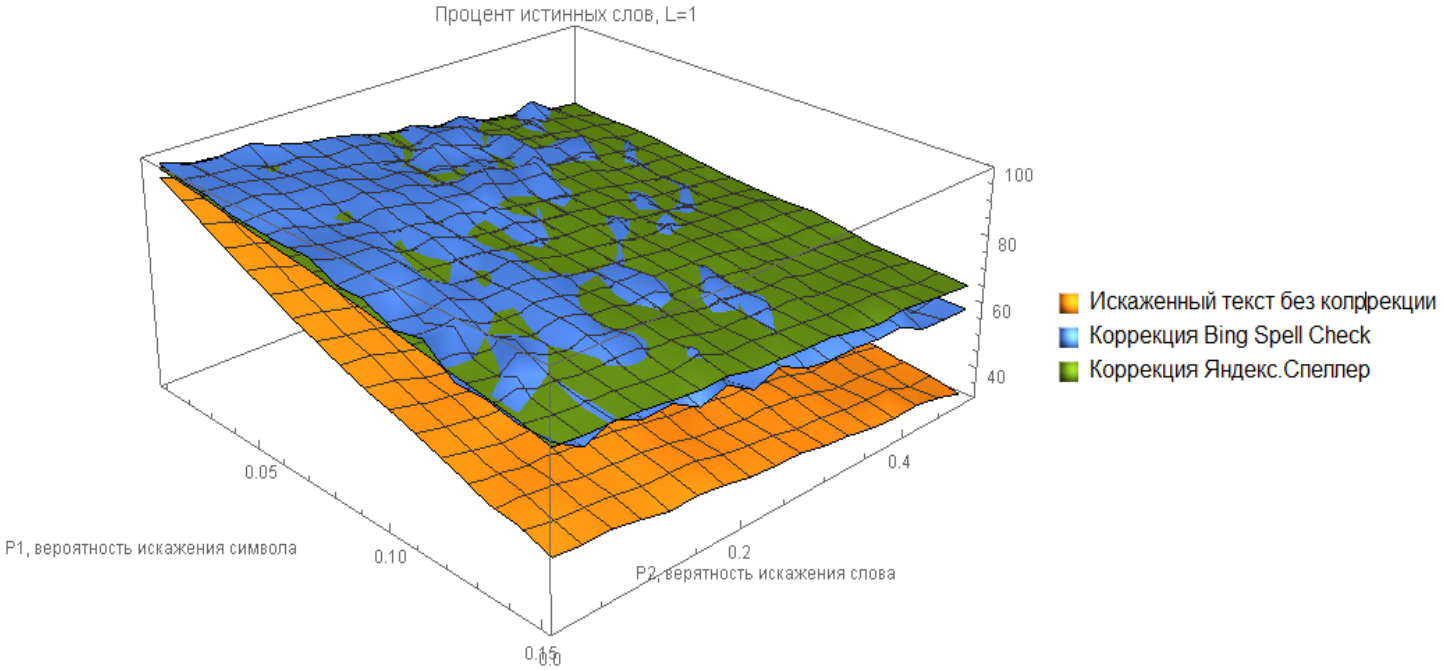


Al Sharou K., Li Z., Specia L. Towards a better understanding of noise in natural language processing , 2021.

ДВУХПАРАМЕТРИЧЕСКАЯ СХЕМА ИСКАЖЕНИЙ. РЕЗУЛЬТАТЫ КОРРЕКЦИИ ОТ BING И ЯНДЕКС

P1 – вероятность замены слов на близкие по Левенштейну, P2 – вероятность посимвольных искажения, L – расстояние по Левенштейну

Область параметров модели искажений: **0.01 <P2< 0.9; 0.01 <P1< 0.5; L=1,2.**



Поверхности Noisy(P1,P2), Yand(P1,P2,1), Bing(P1,P2,1).

Д.А. Бирин, С. Ю. Мельников, В. А. Пересыпкин Об эффективности средств коррекции искаженных текстов в зависимости от характера искажений, 2018

ПРИМЕРЫ ЗАШУМЛЕННЫХ ТЕКСТОВ. СИЛЬНЫЕ, СРЕДНИЕ И СЛАБЫЕ ИСКАЖЕНИЯ

8eb681f801374166a5ce8306b0a43ea4_4.txt — Блокнот

Файл Правка Формат Вид Справка

Мскве. 4 бая. INTERFAX.RU - Итатльясре ВМС првлозят Еврпнеельмик опжерациеюно дставканиыцпушу йптримхеурною 5,7 ты. неыъгаълом, спщахпнымйуаза япоследеесдвое оуырщк а Сревиземно чора, сообаемц агентствбEFE.р в честнст, фегапте "Берсальеё"у ужяй щыосйвлв сорг Рляжо-ди-Клабырия778кмиъратто. Ещек870й жчеловека хрифылих ншаз Сюилияя на сборуъкораблае берхговый охрнн"Аско-29". Кроге кяоио, ы итальянсвий поёрть посмавлены тжъщбе энискоклько ммене мноочиблеунгыём фиупепа нелалм. оВ 2015ъмводу наблюдаемвся ролу поточки нлеральые жбдгрантош, пыающийёсн вохбратаььья пи морвчз Северн Афмикэ к Еровпы. Блолфшинство из цш отйклываюизэ Лкиоиув нкправлениял Иалиц.

8eb681f801374166a5ce8306b0a43ea4_4.txt — Блокнот

Файл Правка Формат Вид Справка

Москва. 4 мая. INTERFAX.RU - Итаальянские ВМС прооди в понедельлникоперацию по доставкецга суму примерна 5,7 тыс. ьелегалов, спасенных за ппследние двое скток в Средиземном мре,сообщхает агетство EFE. В частноти, фрегат"Берсгальекр" уже доставит во порт Реджо-ди-Калабрия 778 мигрантов. Еще 870 человеку прибыли на Сицилию нибоярту корабля хлересголой охтаншу "Ассо-29". Кроме того, в итальянские порты доставлеыы еще несколко менте многочисаленнх гхупп неёегалов. В 2015 году нлблюдается роста потока нелкегальных мигрантом, пытающихся дбраься по молю из Северной Африки в Европу. Большинство и тих отплыва виз Лидии в напраленияйталии.

8eb681f801374166a5ce8306b0a43ea4_4.txt — Блокнот

Файл Правка Формат Вид Справка

Москва. 4 мая. INTERFAX.RU - Итальянские ВМС ыроводетйв понмдельник операцию по доставке на сушу примено 5,7 тыб. нелегалов, спасенных за последние двое суток в Средиземном море, ёсообщает агентстве EFE. В частности, фрегат "Берсальер" уже доставил в порт Реджо-ди-Калабрия 778мигрантов. Еще 870 человек прибыли на Сщилию на борту корабля берегаовой охранф "Ассо-29". Кроье того, в итальянские порты доставлены еще несколько менее многочисленных группу нелегалов. йВ 2015 году наблюдается рост потока нелегальных мигрантов, пытающихся добаться по морю уз Северной Афярики вЕвропу. Большинство виз них отпйыват из Ливии в направлении Италии.

Даже слабые искажения делают невозможным автоматическую обработку текста.

Kumar A., Makhija P., Gupta A. Noisy Text Data: Achilles' Heel of BERT, 2020.

ТЕОРЕТИКО-ИНФОРМАЦИОННЫЙ ПОДХОД. ВОССТАНОВЛЕНИЕ ТЕКСТА КАК ОПТИМИЗАЦИОННАЯ ЗАДАЧА

Пусть канал связи характеризуется набором условных вероятностей $P(S/X)$ появления последовательности наблюдений $S=(s_1, s_2, ..., s_n)$ при условии, что истинная последовательность есть $X=(x_1, x_2, ..., x_n)$, $n=1,2,...$. Задача коррекции состоит в нахождении такой последовательности $\hat{X}=(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n)$ символов из алфавита, для которой достигается максимум условной вероятности $P(X/S)$:

$$\hat{X} = \arg\max_X P(X/S).$$

По формуле Байеса:

$$\arg\max_X P(X/S) = \arg\max_X P(X) \times P(S/X).$$

Таким образом, для коррекции текста необходимо решить оптимизационную задачу

Модель языка $P(X) \times P(S/X) \to \max.$ *Модель искажений*

Величина $P(X)$ определяется моделью языка и обычно представляется в виде

$$P(X) = \prod_{i=1}^{n} p\left(x_i / x_{1,i-1}\right),$$

где $p\left(x_i / x_{1,i-1}\right)$ – вероятность появления символа x_i после фрагмента $x_1, x_2, ..., x_{i-1}$, $i=1,2,...,n$.

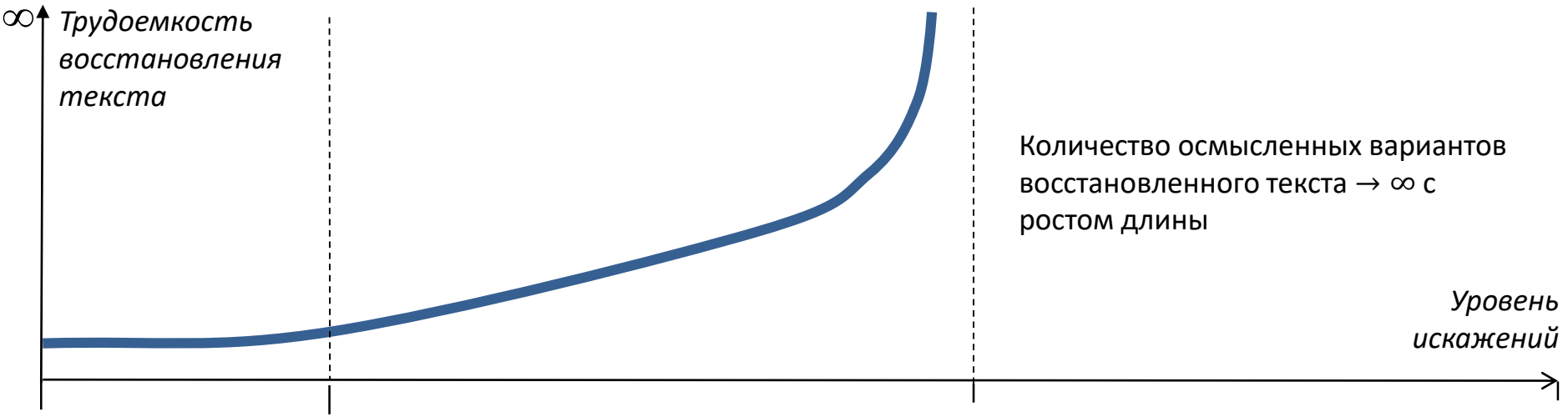
Величина $P(S/X)$ определяется способом искажения текста, ее вид зависит от предметной области, в которой решается задача коррекции.

МЕТОДЫ ВОССТАНОВЛЕНИЯ В ЗАВИСИМОСТИ ОТ УРОВНЯ ИСКАЖЕНИЙ

Методы

Коррекция отдельных ошибок по словарю	Восстановление текста методами поиска с использованием моделей языка	Восстановление текста невозможно
Распространенные спелл-чекеры	Специализированные системы	???

Трудоемкость



ЛОКАЛЬНЫЕ ФЛУКТУАЦИИ ЭНТРОПИИ ТЕКСТА СДВИГАЮТ ГРАНИЦУ «НЕВОЗМОЖНОСТИ ВОССТАНОВЛЕНИЯ»

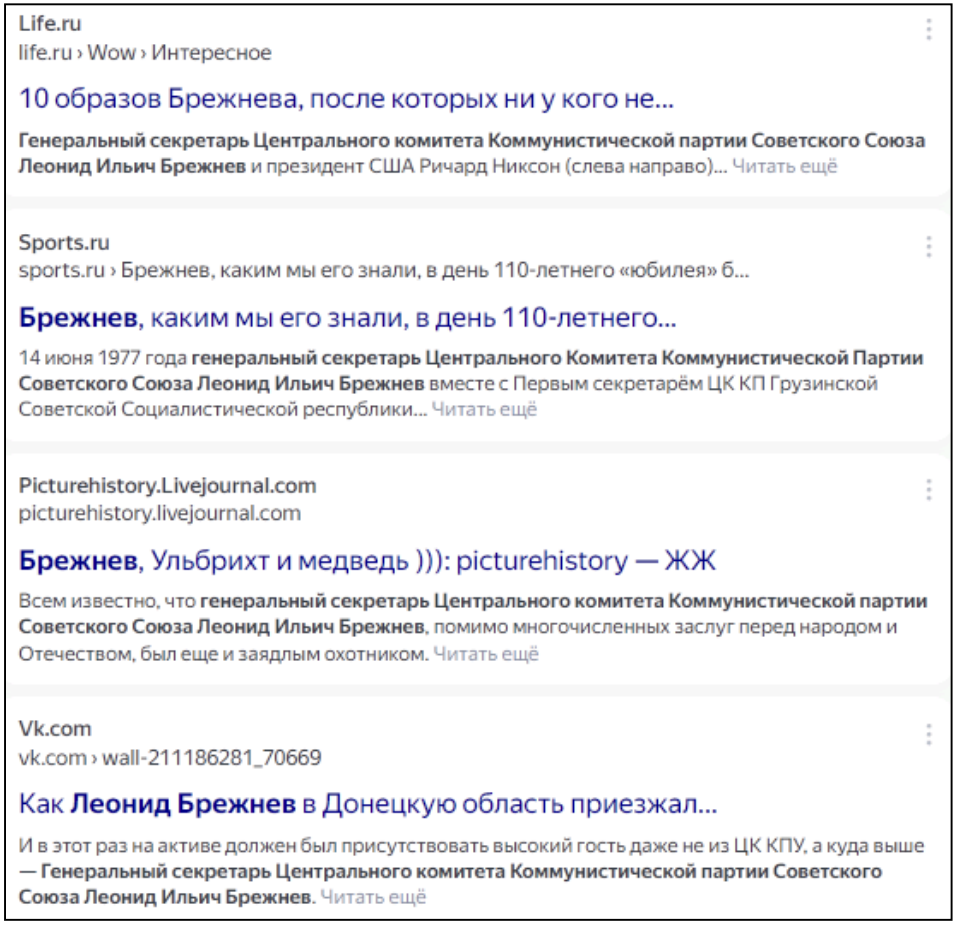
В текстах, в зависимости от их жанра и стиля, могут встречаться низкоэнтропийные участки.

Примеры:

«В целях исключения открытого опубликования материалов, содержащих сведения, составляющих государственную тайну, и сведений, подпадающих под действие контрольных списков по экспортному контролю, в соответствии с федеральными законами от 21.07.1993 № 5485-1 «О государственной тайне», от 18.07.1999 № 183-ФЗ «Об экспортном контроле» и во исполнение Рекомендаций по проведению экспертизы материалов, предназначенных к открытому опубликованию, одобренных Межведомственной комиссией по защите государственной тайны от 24.01.2012 № 225»…(Приказ Тольяттинского государственного университета № 4409 от 10.12.14г.)».

«… и лично генеральный секретарь Центрального комитета Коммунистической партии Советского Союза Леонид Ильич Брежнев»

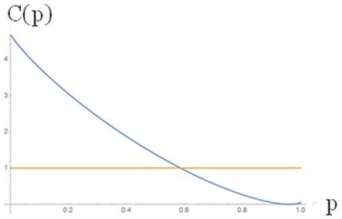
Наличие подобных участков приводит к локальному снижению энтропии текста и может значительно сдвинуть границу «невозможности восстановления».

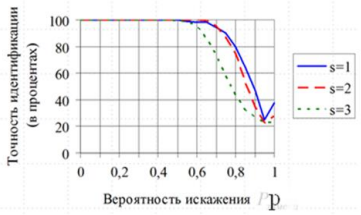


ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ ОЦЕНОК ПРИЗНАКОВ ТЕКСТА В ЗОНЕ «НЕВОЗМОЖНОСТИ ВОСТАНОВЛЕНИЯ»

Такие признаки текста, как язык, тематика, стиль, жанр, авторская принадлежность, и др. могут быть установлены статистически на основе распределения символьных N-грамм.

Тексты на 4-х европейских языках приводились к латинскому алфавиту без пробелов и знаков препинания. Задавалось значение $p \in (0,1)$ вероятности замены символа. Замены осуществлялись независимо друг от друга. Если решение о замене символа принималось, то он заменялся на произвольный другой символ латинского алфавита по равновероятной схеме. Построен дискретный канал без памяти, который моделирует описанные искажения. Вычислена его пропускная способность $C(p)$.





Построена процедура распознавания языка рассматриваемых текстов, использующая s-граммные статистики на символах, $s = 1,2,3$. Согласно основному теоретико-информационному неравенству, исходный текст можно восстановить при условии $H < C$, где $H \approx 1$ - энтропия языка исходного текста, а C - пропускная способность. Для построенного канала при $p < 0.59$ можно восстанавливать исходный текст по искаженному, а при $p > 0.59$ однозначное восстановление текста невозможно. Результаты экспериментов показывают, что язык текста можно определять с высокой степенью надежности даже при $p > 0.59$.

Поэтому, даже в случае, когда восстановить текст нельзя, его содержание можно как-то охарактеризовать.

Кулай А.Ю., Мельников С.Ю. О точности идентификации языка искаженного текста в зависимости от степени искажения, 2009

ОПЫТЫ ПО ПОНИМАНИЮ РЕЧИ В СИЛЬНЫХ ШУМАХ

В период 2010-2019 гг в МГЛУ проводилась серия исследований по восприятию речи в условиях сильных шумов.

Ю. В. Абрамов, Р. К. Потапова, М. В. Хитина. Специфика смыслового восприятия спонтанного звучащего текста (в условиях эскалации уровня шума) (2011)

Хитина М. В. Экспериментальное исследование оценки качества смыслового восприятия спонтанного монолога на фоне аддитивных шумов (2012)

Хитина М. В. Перцептивно-слуховое восприятие материалов разного вида: экспериментальные исследования (2017)

Хитина М. В. Особенности восприятия читаемых текстов в "белом шуме" (2019) и др.

В исследованиях принимало большое количество экспертов (студентов-филологов).

Исследовался, в том числе, вопрос, может ли эксперт воспринять основную и побочные темы произнесенного речевого фрагмента.

Одним из результатов этих исследований был вывод о сильном влиянии квалификации, опыта и индивидуальных качеств воспринимающего речь эксперта и, как следствие, значительном разбросе получаемых оценок.

ВЫВОДЫ

1. Искажения в текстах и зашумление речевого сигнала существенно затрудняют их восприятие человеком и делают практически невозможной автоматическую обработку.

2. Избыточность естественного языка позволяет восстанавливать искаженные тексты, вплоть до уровня шумов, соответствующего шенноновской теоретико-информационной границе. Если уровень искажений превышает эту границу, однозначное восстановление текста невозможно.

3. Для случая, когда уровень искажений в тексте превышает теоретико-информационную границу, предложено два подхода к получению информации из искаженного текста. Первый связан с поиском низкоэнтропийных участков в тексте и их восстановлением, второй связан с возможностью получения интегральной информации о тексте (язык, тематика, жанр и др.) по статистике встречаемости символьных N-грамм.

4. Разработка этих подходов позволит уточнить требования и методики по защите информации для ряда каналов связи.

СПАСИБО ЗА ВНИМАНИЕ

Спасибо за внимание!

melnikov@linfotech.ru