ЗАО «МНИТИ» (г. Москва)

Калегин Сергей Николаевич

Экспериментальное исследование возможности автоматизации процесса языковой идентификации текста

Доклад представлен в 2-х частях:

1. Основные принципы и способы языковой идентификации текста.

2. Сравнение результатов языковой идентификации в автоматическом режиме.

Используемые термины и определения

- Языковая идентификация и определение языка текста
- Определение языка и языковой принадлежности текста

Цель исследования

Определить возможность полной автоматизации процесса языковой идентификации неструктурированного текста на основе известных решений.

Задачи исследования

1. Определить эффективность используемых способов языковой идентификации текста.

2. Выяснить вероятность верной идентификации языковой принадлежности текста текста с помощью существующих программных решений.

Актуальность исследования

В условиях техногенного информационного общества при постоянно увеличивающихся потоках информации обилие различных интеллектуальных систем обработки неструктурированных данных, к которым относятся программные комплексы по сбору и обработке информации, системы обработки корреспонденции, каталогизации и сортировки литературы, машинного перевода и информационно-поисковые системы, базы знаний, системы глобального мониторинга и т.д., требует автоматизации процесса идентификации языка до начала обработки данных. Решение подобной задачи традиционными способами (вручную или полуавтоматически) не позволяет исключить человеческий фактор на этапе подготовки данных к обработке, а также полностью автоматизировать подобные системы, что отрицательно влияет на их производительность и затормаживает развитие. Поэтому проведённое исследование является актуальным как в теоретическом, так и практическом плане.

Практическое применение в сфере «больших данных»

- Языковая идентификация – первый этап обработки любого неструктурированного многоязычного массива данных.
- В системах глобального мониторинга без идентификации языковой принадлежности информации обработка данных практически невозможна.
- Смысловой поиск информации в мировом масштабе без её языковой идентификации не представляется возможным.
- Тематический анализ информации без языковой идентификации невозможен.
- Автоматическое пополнение, обновление и расширение глобальных баз данных невозможно без языковой идентификации добавляемой информации.

Часть 1. **Основные принципы и способы языковой идентификации текста**

Основные принципы языковой идентификации:

- символьный;
- лексический;
- грамматический;
- лингвистический.

Основные способы языковой идентификации:

- использование словарей;
- поиск коротких слов;
- использование характерных символов и сочетаний;
- использование статистики комбинаций символов (n-грамм);
- морфологический анализ слов;
- грамматический анализ предложений;
- лингвистический анализ предложений;
- выделение служебных слов (артиклей и т.п.).

Символьные способы идентификации

- Использование характерных символов и/или их сочетаний.
- Использование статистики комбинаций символов (n-грамм) и языковых моделей.

Характерные символы

Кириллические: Ё, Й, Ъ, Ь, Ғ, Ҙ, Ҥ, Ҷ, Ӌ, Ҟ …
Латинские: Ł, À, Æ, Ê, Ï, Ü, Ä, Ö, Õ, Ñ, Í, Č, Ď, Ħ, Ş, Ŝ, Ž …

Характерные сочетания символов

SCH, TSCH, CK, CZ, DZ, DŹ, DŻ, RZ, SZ, SZCZ, ŚĆ, ŹDŹ, ŻDŻ …

Таблица n-грамм

Шведский	Английский	Немецкий	Французский	Итальянский
en_	_th	en_	_de	_di
_._	he_	er_	es_	to_
er_	the	_de	de_	_de
et_	_,_	der	ent	di_
tt_	nd_	ie_	nt_	_co
_de	ed_	ich	_le	la_
ar_	_an	sch	e_d	re_
-,-	and	ein	le_	ion

Примеры идентификации с помощью поиска характерных символов

Языковая идентификация текстов на русском (славянская группа) и кумыкском (тюркская группа) языках по 4 характерным символам: **ё, й, ъ** и **ь.**

Русский язык

*В Курске возбудили дело против депутата областно**й** думы Ольги Ли, которая разместила в интернете видеоролик с обращением к президенту Владимиру Путину. В этих выступлениях следователи нашли возбуждение ненависти к власти….*

*Представитель Минобороны ДНР Эдуард Басурин проинформировал, что сегодня была совершена очередная попытка прорыва обороны ополчения республики на линии соприкосновения в пригороде Донецка. По словам представителя Минобороны, силовики ВСУ пытаются реализовать**ь** свои предыдущие планы по разъединению стратегически значимых населенных пунктов ДНР.*
(Отрывки из новостной ленты)

*…Наша способност**ь** усваивать необычное не беспредельна, а когда путешествуешь на другие планеты, пределы оказываются очен**ь** узкими. Слишком много новых впечатлений; их приток становится невыносимым, и мозг ищет отдыха в буферном процессе аналогизирования. Этот процесс как бы созда**ё**т мост между воспринятым известным и неприемлимым неизвестным, облекает невыносимое неизвестное в желанную мантию привычного….* *(Роберт Шекли «Обмен разумов»)*

Всего символов в текстах на русском языке **870.**
Из них **5** характерных, что составляет ~ **0,6 %** от общего количества знаков.

Кумыкский язык

…Сююв юрек яллатгъансан
Мени ойлягъа салгъансан
Гёзлеримни пашмалыкъ,
Гёз яшларым акътыргъансан
Сюювгъе инандырдип ,
Сюювну алмайдип,
Сенсиз яшап болмай эдим,
Сюедим сени, сюедим
Айтдим мен сагъа, сюйгеним,
Унутма мени, тилеймен,
Унутдунгму сен, аявлум,
Башхагъа гёзюнг гъарайму?

Англамадынг сен мени,
Мени сен англамадынг.
Мен айтагъан сёзлеге бурулуп сен гъарамадынг ….

(Отрывок из песни «Сююв юрек яллатгъансан, сюйгеним»)

Всего символов в тексте на кумыкском языке **320.**
Из них **22** характерных, что составляет ~ **6 %** от общего количества знаков.

Подобное сравнение с текстом на немецком языке, характерным признаком которого являются буквы «**ä**» «**ö**» и «**ü**».

Немецкий язык

„Wir nehmen im Match diesen Ball auf und werden ihn weiter voranspielen“, sagte Schäubles Sprecher in Berlin. Die sogenannten Panama Papers seien keine Überraschung, erhöhten aber den Druck auf Steueroasen auf der ganzen Welt. Das „Unterholz“ bei Versuchen, die Steuerbehörden auszutricksen, müsse besser ausgeleuchtet werden. Schäuble selbst werde vor der Frühjahrstagung des Internationalen Währungsfonds (IWF) und der Weltbank vom 15. bis zum 17. April in Washington die Initiative ergreifen in der Frage, wie es international mehr Transparenz gegen illegale Finanzgeschäfte geben könne, kündigte Jäger an. *(Из новостной ленты)*

Содержит ~ **2,3 %** характерных символов.

Турецкий язык

Uzun yıllar önce Çin'de bir kral vardı. Kralın sarayı çok büyük ve çok güzeldi. Çatısı altındı. Pencerelerinde bin tane lâmba vardı. Koridorları uzun ve bahçeleri sayısızdı.
Sarayın çevresinde yeşil bir orman ve mavi bir deniz vardı.
Ormanda sayısız hayvan vardı. Fakat hayvanların en meşhuru küçük gri bülbüldü. Sesi çok güzel ve harikaydi.
İnsanlar her yerde bülbülün güzel şarkılarından bahsetti.
Balıkçılar deniz kenarında bülbülün güzel sesini dinledi. Herkes bülbülün güzel şarkılarını duydu ama kimse onu görmedi.
Bülbül Çin'de ve komşu ülkelerde meşhur oldu. Uzak ülkelerden insanlar bülbülü dinlemek için ormana geldiler. Şairler bülbül için şiirler yazdılar. Ülkede herkes bülbülün ününü duydu. Yalnız ülkenin kralı bundan haberdar değildi. *(Отрывок из сказки Kral ve bülbül)*

Содержит ~ **6,2 %** характерных символов.

Вывод. Результаты анализа текстов показывают несостоятельность символьного способа.

Лексические способы идентификации

- Использование словарей всех определяемых языков.
- Поиск характерных коротких слов в тексте.

Анализ текстов словарным способом

1. ***Представитель*** *Минобороны ДНР Эдуард Басурин проинформировал, что* ***сегодня*** *была совершена очередная* ***попытка*** *прорыва обороны ополчения республики на линии соприкосновения в пригороде Донецка. По словам представителя Минобороны, силовики ВСУ пытаются* ***реализовать*** *свои предыдущие планы по разъединению* ***стратегически*** *значимых населенных пунктов ДНР.*
(Из новостной ленты)

2. *Нова была красивой планетой, первой успешной земной колонией.* ***Сейчас*** *это* ***пустыня****. Целые города исчезли с её лица, уничтоженные взрывами нейтронных бомб. Нечего* ***опознать****. Нечего* ***похоронить****. Некого* ***оплакать****.* ***Вторжение*** *началось* ***внезапно****. Объединённые силы Земли нанесли коварные удары по всей территории планеты.* ***После*** *них остались растерзанные тела. Крики женщин и детей о помощи захлебнулись в плазменном огне, прожегшем их* ***плоть****. Мы слышали* ***много*** *проповедей о добродетелях прогресса и науки. Что хорошего от них, если целые цивилизации разрушаются в* ***мгновение*** *ока?*
(Вступление из компьютерной игры "Power DOLLS")

В первом тексте из 40 слов 5 стоят в словарной форме и имеют длину более 4 символов, что составляет **12,5 %** от общего числа.
Во втором тексте содержится ~ **14 %** словарных форм из 78 слов (в среднем каждое седьмое слово), остальные формы при таком подходе в идентификации не участвуют.

Вывод. Применение данного способа для коротких текстов нецелесообразно, так как пропускаются целые предложения.

Языковая идентификация текстов по служебным и коротким словам

Польский язык

Dawno, dawno temu wędrowali trzej bracia – Lech, Czech ***i*** *Rus.*
Jechali już bardzo długo, każdy z nich był przywódcą dzielnych wojów ***i*** *każdy miał pod opieką swój lud.* ***Na*** *wozach jechały kobiety z dziećmi, ciągnięto cały dobytek, stada bydła. Wszyscy byli już bardzo zmęczeni* ***i*** *szukali miejsca* ***do*** *odpoczynku* ***a*** *najchętniej miejsca, w którym mogliby się osiedlić, czyli zamieszkać* ***na*** *stałe.*

Nagle wjechali ***na*** *dużą polanę,* ***na*** *tej polanie rósł olbrzymi dąb,* ***a na*** *tym dębie białe orły miały swoje gniazdo.* ***To*** *miejsce bardzo spodobało* ***się*** Lechowi, *nigdy przed tym nie widział tak pięknego miejsca. Postanowił, że wraz* ***ze*** *swoim ludem* ***tu*** *pozostanie. Gdy Lech patrzył* ***na*** *orła,* ***ten*** *nagle rozpostarł skrzydła* ***na*** *tle nieba czerwonego od zachodzącego słońca. Postanowił Lech, że odtąd taki będzie znak jego narodu.*
(*Легенда про Леха, Чеха и Руса.****)***

Przewodniczący Rady Europejskiej opowiadał też, że podczas jego rozmów z innymi premierami ***i*** *prezydentami największe zainteresowanie budziły dwie sprawy. - Byłem trochę tym zaskoczony, ale* ***po*** *chwili zrozumiałem dlaczego.* ***To*** *były pytania* ***o*** *Puszczę Białowieską* ***i o*** *stadninę koni. Może dlatego, że* ***to ma*** *też wymiar symboliczny. Muszę powiedzieć, że dla mnie jako Polaka ktoś, kto wycina starodrzew albo doprowadza* ***do*** *śmierci koni -* ***i to*** *w takiej stadninie jak Janów - robi straszne rzeczy. Zastanawiam* ***się****, czy nie zaczną też strzelać* ***do*** *bocianów – powiedział.*
(Отрывок из газеты)

Чешский язык

*Byli jeden dědeček **a** babička **a** měli malou vnučku. Každý **den** vnučka Evička pomáhala svým prarodičům. Evička krmila zvířátka, babička pracovala v kuchyni, děda měl **na** starost pole. Vypěstoval řepu takovou, že **se na ni** chodili dívat sousedi. Byla veliká, zabírala polovinu pole **a** stále rostla **a** rostla. Co dělat? Děda ji musí vykopat. Motyka **se** mu zlomila. Co teď? Jak **tu** řepu dostat?*
*Zavolal **na** pomoc babičku. Ta vzala dědu v pase, **ten** chytil **ze** všech sil silné listy **a** společnými silami táhli, táhli, ale řepu nevytáhli.*
(Отрывок из сказки про репку)

Английский язык

*Once there was **a** Prince who wanted **to** marry **a** Princess. Only **a** real one would **do**. **So he** traveled through all the world **to** find **her**, and everywhere things went wrong. There were Princesses aplenty, but how **was he to** know whether they were real Princesses? There **was** something not quite right about them all. **So he** came home again and **was** unhappy, because **he** did **so** want **to** have **a** real Princess.*
(Отрывок из сказки The Princess on the Pea)

*"**Ten** Years exposed the fear **of** Hong Kong people (towards China)," said one **of** the film's directors, Chow Kwun-wai.*
*Producer Andrew Choi told the BBC the award came **as a** surprise.*
*"It **is** important for Hong Kong that **a** film that echoes **so** much **of** what people are feeling **in** their hearts **has** won."*
(Из новостной ленты)

Немецкий язык

„Wir nehmen im Match diesen Ball auf und werden ihn weiter voranspielen“, sagte Schäubles Sprecher ***in*** *Berlin. Die sogenannten Panama Papers seien keine Überraschung, erhöhten aber* ***den*** *Druck auf Steueroasen auf* ***der*** *ganzen Welt.* ***Das*** *„Unterholz“ bei Versuchen,* ***die*** *Steuerbehörden auszutricksen, müsse besser ausgeleuchtet werden. Schäuble selbst werde vor* ***der*** *Frühjahrstagung* ***des*** *Internationalen Währungsfonds (IWF) und* ***der*** *Weltbank vom 15. bis zum 17. April* ***in*** *Washington* ***die*** *Initiative ergreifen* ***in der*** *Frage,* ***wie es*** *international mehr Transparenz gegen illegale Finanzgeschäfte geben könne, kündigte Jäger* ***an****.*

„Wir müssen Briefkastenfirmen und Stiftungen, deren wirtschaftlich Berechtigte anonym bleiben, weltweit verbieten", sagte ***der*** *SPD-Politiker* ***der*** *„Süddeutschen Zeitung“.* ***Er*** *sprach von „Geldgier* ***der*** *Superreichen“,* ***die*** *sich verbinde „****mit der*** *Gewissenlosigkeit* ***im*** *Banken- und Finanzsektor“. Beides zerstöre* ***das*** *Vertrauen* ***in den*** *Rechtsstaat.* ***Es*** *gehe* ***um*** *„organisierte Kriminalität von Banken und Finanzjongleuren“,* ***die mit*** *allen Mitteln zu bekämpfen* ***sei****.* *(Из новостной ленты)*

Голландский язык

De *taal* ***en het*** *literaire leven* ***in de*** *steden* ***en*** *dorpen van Nederland* ***en*** *Vlaanderen. Wie werd waar geboren? Welke romans spelen* ***in*** *Antwerpen, hoe klonk het Katwijks* ***in de*** *19de eeuw?* ***De*** *Atlas* ***is*** *een snel groeiende informatiebron voor plaatsgebonden verschijnselen* ***in de*** *Nederlandstalige cultuur. Alfabetische overzichten* ***en*** *aanklikbare kaarten wijzen* ***de*** *weg.* ***En wie*** *alleen* ***in de*** *Middeleeuwen* ***of in de*** *Achttiende Eeuw wil rondkijken,* ***kan*** *dat doen aan* ***de*** *hand van deeloverzichten* ***en*** *historisch kaartmateriaal. Daarnaast* ***is er*** *een afdeling Buitengaats, waar alle mogelijke literatuur* ***is te*** *vinden* ***over de*** *relaties van Nederlandstaligen* ***met de*** *rest van* ***de*** *wereld.*
(Из новостной ленты)

Сводная таблица совпадений коротких слов в разных языках

№ п/п	Идентифицирующее слово	В каких языках встречается
1	a	польский, чешский, английский, кастильский (испанский), португальский, румынский, французский, гаэльский, идо
2	na	польский, чешский, гаэльский
3	do	польский, английский, португальский, гаэльский
4	to	польский, английский
5	i	польский, шведский, гаэльский, итальянский
6	ten	польский, чешский, английский
7	o	польский, португальский, румынский
8	se	чешский, кастильский, португальский, французский
9	den	чешский, немецкий, шведский
10	of, is	английский, голландский
11	as	английский, португальский, гаэльский
12	in	английский, немецкий, голландский, итальянский, гаэльский
13	das	немецкий, португальский
14	er	немецкий, голландский
15	es	немецкий, кастильский
16	an	немецкий, гаэльский, идо
17	sei	немецкий, итальянский
18	kan	голландский, шведский
19	over	голландский, румынский
20	en	голландский, кастильский, французский, эсперанто, идо
21	de	голландский, шведский, итальянский, кастильский, португальский, румынский, французский, эсперанто

Характерные особенности основных способов

В таблице представлены способы языковой идентификации (в порядке уменьшения популярности) с указанием их преимуществ и недостатков.

№ п/п	Способ	Идентифицирующие элементы	Преимущества	Недостатки
1	Словарный	слова в исходной форме	простота реализации	низкая эффективность и повышенная ресурсоёмкость
2	Символьный	характерные символы традиционной письменности	простота реализации	низкая эффективность и высокая вероятность ошибок
3	Сравнение n-грамм-моделей	характерные сочетания символов	высокая эффективность при определённых условиях	сложность реализации, повышенная ресурсоёмкость и вероятностный результат
4	Поиск коротких слов	характерные короткие слова (до 4-5 символов)	простота реализации	низкая эффективность и высокая вероятность ошибок
5	Грамматический анализ	характерные аффиксы, словоформы и грамматические особенности	высокая эффективность	требуется подготовительный этап и грамматические анализаторы для каждого языка, сложность реализации и высокая ресурсоёмкость

Часть 2. **Сравнение результатов языковой идентификации в автоматическом режиме**

Тестируемые программы

- **Guesser** – автоматический определитель языка текста российской компании «Фларус». Идентифицирует язык текста при помощи словарей и характерных символов или их сочетаний.
- **Automatic language identifier** – программа-определитель итальянского исследовательского центра Translated Labs. Определяет язык по списку сочетаний символов – n-грамм, представляющему собой своеобразную модель конкретного языка.
- **TextCat** – один из первых определителей на базе n-грамм-моделей.
- **Полиглот 3000** – программа кипрской компании Likasoft. Также использует n-граммы, как и две следующих системы.
- **Language Identifier by Henrik Falck** – программа Генриха Фалка.
- **SILC** – Système d'Identification de la Langue et du Codage. Система определения языка и кодировки текста от канадской лаборатории RALI.
- **Xerox Language Identifier** – гибридный языковой определитель от американской компании Xerox. Комбинирует несколько способов идентификации.

Этапы тестирования

На **1-м этапе** всеми программами производился анализ отдельных предложений на языках различных генеалогических групп Европы. Предложения составлены в соответствии с требованиями применяемых способов идентификации.

На **2-м этапе** тем же программам передавались тексты объёмом в 1 абзац (30-80 слов) на языках всех генеалогических групп данного региона: кельтской, романской, германской, славянской, балтийской, финно-угорской, греческой и палеобалканской, а также на искусственных языках: воляпюк, эсперанто, идо. Это позволило определить зависимость верности идентификации от количества языков и объёма текста. Условия тестирования и анализируемые тексты для всех программ были одинаковые, ресурсоёмкость не учитывалась.

Результаты 1-го этапа

Примеры анализа фраз на 3-х различных европейских языках: **итальянском** (западнороманской подгруппы), **немецком** (западногерманской подгруппы) и **польском** (западнославянской подгруппы).

На итальянском языке.

Фраза 1: "**Io non capisco una parola di ciò che dice.**" (*Я не понимаю ни слова в том, что он говорит*). Содержит 9 слов, 42 символа.

Фраза 2: "**Molte cose resteranno le stesse ma tante altre cambieranno**" (*Многие вещи остаются такими же, но многие другие изменяются*). Содержит 9 слов, 58 символов.

На немецком языке.

Фраза 1: "**Was kann ich heute machen**" (*Что я должен сегодня делать*). Содержит 5 слов, 25 символов.

Фраза 2: "**Das neue Solutions Centre bietet ein breites Portfolio an technischen Dienstleistungen und modernen Schulungseinrichtungen**" (*Новый солюшен-центр предлагает широкий спектр технических услуг и современных средств обучения*). Содержит 14 слов, 122 символа.

На польском языке.

Фраза 1: "**O swoich dalszych krokach adwokat mówić na razie nie chce**".
(*О своих дальнейших шагах адвокат говорить пока не хочет*). Содержит 10 слов, 57 символов.

Фраза 2: "**Nowy Rok to najlepsza okazja do zmiany swojego nastawienia do ludzi**" (*Новый год – это лучшая возможность изменить своё отношение к людям*). Содержит 11 слов, 67 символов.

Итальянский язык

Фраза 1: "**Io non capisco una parola di ciò che dice.**"

№ п/п	Предполагаемый язык текста	Вероятность соответствия
1	португальский	10 %
2	мальтийский	2 %
3	финский	8 %
4	немецкий	8 %
5	нидерландский	8 %
6	эстонский	8 %
7	словацкий	8 %
8	венгерский	8 %
9	шведский	8 %
10	испанский	8 %
11	турецкий	8 %
12	норвежский	8 %
13	датский	8 %
14	итальянский	2 %

Тестирование других программ

Фраза 2: "**Molte cose resteranno le stesse ma tante altre cambieranno**"

№ п/п	Программа идентификации языка	Результат
1	Automatic language identifier (T-Labs)	**Samoan**
2	Полиглот 3000	Не распознан
3	TextCat	italian
4	Language Identifier by Henrik Falck	Italian (or possibly Spanish)
5	SILC	Italian

Результаты 2-го этапа

Русский язык

Строительство железнодорожных подходов будет синхронизировано со строительством железнодорожного моста. Их сдадут заблаговременно до открытия железнодорожного моста», — подчеркнули в министерстве.

Guesser: русский (ы) 61 %, болгарский 38 %;
T-Labs language identifier: Bulgarian;
Полиглот 3000: Русский 65 %;
TextCat: turkish;
Henrik Falck identifier: Russian;
SILC: Russian;
Xerox language identifier: Русский язык (Russian).

Сербский язык

Био сам још сасвим млад послушник кад сам почео. Отац Амвросије ми је рекао: "Што год радиш, непрестано говори у себи: "Иисусе, сине Божији, помилуј ме!" Био сам дечак и свим срцем га послушао. Сваког дана бих исповедао оцу духовнику шта се унутри у души догађало, а он је саветовао шта да чиним.

Guesser: русский 52 %, болгарский 47 %;
T-Labs language identifier: Bulgarian;
Полиглот 3000: Сербский 73 %;
TextCat: nepali;
Henrik Falck identifier: Serbian;
SILC: Russian;
Xerox language identifier: Српски (Serbian).

Латышский язык

Tāpēc viņš no jauna izņēma no kastes tējas katliņu, kuram bija pazudusi āpša galva un aste, un kurš bija atguvis savu agrāko izskatu, un parādīja to skārdniekam. Apskatījis tējas katliņu, skārdnieks par to piedāvāja divdesmit vara monētas.

Guesser: венгерский (zs) 32 %, эстонский 8 %, нидерландский 7 % и т.д.;
T-Labs language identifier: Latvian;
Полиглот 3000: Латышский 100 %;
TextCat: latvian;
Henrik Falck identifier: Serbian (or possibly Bosnian, Croatian, or Slovenian);
SILC: Serbo-Croatian;
Xerox language identifier: Latviešu valoda (Latvian).

Эстонский язык

Koit Toome sõnul on tal väga hea meel teha taas koostööd Getteri ja Sven Lõhmusega. Ka ühise loo ajastus on suurepärane, kuna neil on Getteriga suvel palju ühiseid esinemisi, mille seas ka juulis ühine suvetuur. See tähendab, et duett kõlab lisaks stuudioversioonile ka läbi suve lugematul arvul kontsertlavadel.

Guesser: эстонский (öö) 29 %, нидерландский 13 %, финский 9 % и т.д.;
T-Labs language identifier: Fijian;
Полиглот 3000: Эстонский 77 %;
TextCat: estonian;
Henrik Falck identifier: Estonian;
SILC: Estonian;
Xerox language identifier: Eesti (Estonian).

Итоги сравнительного исследования

Эффективность способов при идентификации языка неструктурированного текста варьируется в широких пределах – обычно от 3**0 до 100 %** – и зависит от многих факторов, среди которых можно выделить следующие:

- количество определяемых языков;
- генеалогическое родство определяемых языков;
- используемая письменная система анализируемого текста;
- используемая кодировка текста;
- количество слов в анализируемом тексте (объём текста);
- применяемые вариант и алгоритм идентификации;
- качество языковых моделей в базе определителя;
- размер и качество словарей или идентифицирующих матриц.

Например, **при 2 неродственных** определяемых языках и относительно длинном тексте, записанном традиционной для данного языка письменностью в кодировке **Unicode**, вероятность определения может достигать **100 %**. Однако при **увеличении количества языков** и смешении письменных систем **результативность существенно понижается** и может доходить до бесполезных результатов, когда вероятность ассоциации текста с различными языками становится практически одинаковой.

Выводы

1. Результат определения языковой принадлежности существующих решений зависит от их **способов идентификации и технической реализации**, а также объёма анализируемого текста.

2. Все современные способы **зависят от применяемой письменной системы**, а также **соблюдения правил грамматики и орфографии** в анализируемом тексте, что обуславливает невозможность их использования при записи текста, нелитературным, нетрадиционным или несовременным способом.

3. Рассмотренные языковые определители **малопригодны для работы с короткими сообщениями**, которые используются в большинстве современных коммуникационных систем.

4. На сегодняшний день **не существует универсальных способов и технических решений** для гарантированного **автоматического определения** языковой принадлежности неструктурированного текста. Любой способ или языковой определитель имеет свои преимущества, недостатки и особенности применения в каждом конкретном случае.

Заключение

В результате проведённого исследования установлено, что **современные решения** в области машинной идентификации языковой принадлежности текста **не позволяют полностью автоматизировать данный процесс и требуют постоянного контроля человеком**.

Подобное программное обеспечение относительно успешно может применяться при небольшом количестве заранее известных и кардинально отличных неродственных языков в некритических задачах. Например, при сортировке оригинальной русской, арабской и китайской литературы в электронных библиотеках, в системах селекции двуязычных данных и т.д., где ошибки идентификации не приведут к сбою всего программного комплекса.

Доклад закончен.

Благодарю за внимание.