Інтелектуальний аналіз діалектних текстів

Голуб Марія
кафедра інформаційної безпеки та комп'ютерної інженерії
Черкаський державний технологічний університет
Черкаси, Україна
mas-golub@yandex.ua

The results of research processes modeling of dialect texts within the multi- level information monitoring technology are regarded in the article. A new method of classifying text messages at the residence of their authors is proposed. The signs for classification got after decomposition of text and calculating their frequency characteristics. For synthesis models used GMDH. The quantity correctly classified texts from 78% to 100%. Convert text messages in an array input allows the advantage of multi-level modeling techniques in information technology monitoring text messages.

Ключові слова: індуктивне моделювання, аналіз тексту, інформаційний моніторинг.

Вступ

Інтелектуальний аналіз діалектних текстів дозволяє виявити найбільш значимі властивості авторів та відобразити їх в структурі багатопараметричних моделей. Ці моделі розв'язують слабоформалізовані завдання класифікації текстів за властивостями авторів, виконуючи функції вирішуючих правил. Процеси визначення характеристик автора друкованого тексту набувають особливої актуальності в сучасних умовах інформаційної війни.

Найбільш ефективними засобами автоматизації процесів виявлення характеристик авторів друкованих текстів є методики профілювання текстів, в яких використано методи статистичного моделювання. На думку авторів, одним із найбільш вдалих прикладів використання такого підходу є серія робіт Т.А. Литвинової, зокрема робота [1]. Застосування регресійно-кореляційного аналізу дозволило отримати множину моделей, що уможливлюють виявлення статі автора, оцінку рівня самоконтролю, емоційної врівноваженості, практичності. Для виконання кожного завдання необхідно отримати масив чисельних характеристик тексту достатньої інформативності і метод синтезу моделей достатньої потужності.

Моделювання діалектного тексту має на меті виявлення місця проживання його автора. Достатню потужність, для моделювання діалектних текстів, має технологія багаторівневого моніторингу, що реалізована у формі моніторингової інформаційної системи (МІС) [2]. Інформаційні системи цього класу, як правило, забезпечують кілька типів технологій моніторингу, які різняться між собою процесами формування ПО та забезпечення інформативності показникам, що утворюють масив вхідних даних (МВД). На етапі ж синтезу моделей розв'язуються типові завдання ідентифікації функціональних залежностей, класифікації, розпізнавання образів, прогнозування та ін.

При реалізації технології інформаційного моніторингу в МІС інтелектуальний аналіз даних забезпечується переліком алгоритмів синтезу моделей (АСМ), значна кількість яких ґрунтується на індуктивних методах [3]

Мета роботи

Метою статті є розробка нового методу класифікації текстів за говірками їх авторів шляхом поєднання процедур перетворення тексту в масив чисельних характеристик та побудови вирішуючого правила у вигляді багатопараметричної моделі-класифікатора для виконання завдання виявлення місця проживання автора текстового повідомлення.

Таким чином необхідно автоматизувати процес класифікації текстових повідомлень. Це завдання слабоформалізоване, оскільки успішно його виконати за допомогою переліку заданих ознак із однозначно визначеними чисельними характеристиками не вдається.

РЕЗУЛЬТАТИ ДОСЛІДЖЕНЬ

Була сформульована гіпотеза про те, що вирішуюче правило необхідно будувати у вигляді індуктивної моделі за багаторядним алгоритмом МГУА.

Досліджено особливості формування масиву вхідних даних (МВД) та процесу синтезу багатопараметричних моделей [4], здатних класифікувати текстові повідомлення за їх належністю до різних типів говірок, що притаманні населенню центральної, північної, південної, західної та східної частин Середньої Наддніпрянщини.

При формуванні МВД у таблицю поєднані значення частотних характеристик показників тексту, перелік яких поданий у [5]. Частотні характеристики були розраховані на окремих вікнах – ділянках тексту, які містили по 5000 знаків.

У результаті для синтезу моделей використано 119 точок спостережень первинного опису. Їх розбито на послідовності *A* і *B* для формування зовнішнього критерію селекції моделей. Ще 11 точок утворювали послідовність *C*, їх використано для випробувань готових моделей, але у процесі синтезу цих моделей вони участі не брали. У процесі синтезу моделі розв'язано завдання класифікації точок спостереження. Модель навчалась зараховувати тексти із табл. 2, описані точками спостереження в масиві даних ПО, до конкретних класів, поданих у табл. 2. Після навчання моделі отримували назви, що збігаються з населеними пунктами, які були об'єктами для моделювання.

У табл. 1 подані результати випробувань отриманих моделей.

Таблиця 2. Результати випробування моделей

Місце проживання автора	Кількість правильно класифікованих точок спостереження, %
Соболівка	96,43
Ладижинка	92,86
Гельмязів	82,14
Москаленки	85,71
Богодухівка	82,14
Ірклiїв	89,29
Зорівка	100,00
Кононівка	89,29

Результати випробування моделей, подані у табл. 1, свідчать, що вдалось синтезувати корисні моделі, здатні виконувати функції інтелектуального аналізу діалектних текстів.

ЛІТЕРАТУРА

[1] Формально-грамматические корреляты личностных особенностей автора письменного текста / Т.А. Литвинова // Филологические науки. Вопросы теории и практики. – 2013. – № 12 (30), Ч. 1. – С. 132 – 135.

[2] Голуб С.В. Багаторівневе моделювання в технологіях моніторингу оточуючого середовища / Голуб С.В. – Черкаси: Вид. від. ЧНУ імені Богдана Хмельницького, 2007. – 220 с.

[3] Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем / Ивахненко А.Г. – К.: Наукова думка, 1981. – 296 с.

[4] Голуб С.В. Відображення властивостей автора тексту в структурі багатопараметричної моделі / С.В. Голуб, О.В. Константиновська, М.С. Голуб // Системи обробки інформації: зб. наук. праць. – Х.: Харківський університет повітряних сил імені Івана Кожедуба, 2014. – Вип. 9 (125). – С. 82 – 87.

[5] Мартинова Г. Середньонаддніпрянський діалект. Фонологія і фонетика / Мартинова Г. – Черкаси: Тясмин, 2003. – 356 с.