УДК 004.9
П. Кміть, С. Лупенко
(Тернопільський національний технічний університет імені Івана Пулюя)

ВИКОРИСТАННЯ ПРИХОВАНОЇ МАРКОВСЬКОЇ МОДЕЛІ В СИСТЕМАХ ДЕТЕКТУВАННЯ ЕМОТИКОНІВ

В розроблюваній системі детектування смайлів використовується система котра за своїми характеристиками схожа на просту Марковську модель, котра ділить стрічку графем Юнікоду на суміжні ділянки, котрі в основному бувають цілі емотікони, або такі що складаються з декількох частин (знаки пунктуації та інші символи).

Отже, емотіконами, в основному, є дані нелінгвістичного характеру. Підхід забезпечує високу чутливість і повертає майже всі речення у котрих є емотікони, але має низьку точність, тому що може прийняти за смайл ділянки з великою кількістю символів пунктуації. Проста ПММ(прихована марківська модель) складається з двох станів: A(головним чином лінгвістичний) та @ (головним чином нелінгвістичний). Зважаючи на те, що існує два класи результуючих символів ПММ повинен мати дві можливості результату: одну для їхніх основних класів символів(лінгвістичних L і нелінгвістичних N) і одну для іншого класу символів.

Нелінгвістичні символи часто з'являються в лінгвістичних послідовностях. Однак, послідовність з трьох таких символів зустрічається рідко, лінгвістичні символи також часто зустрічаються в емотіконах, але кількістю не більшою за три. Оскільки, для сегмента в суміжних послідовностях з певною кількістю символів в ряд, можливість переходу зі стану A в стан @ має бути набагато нижча, чим можливість результату з однієї чи двох N зі станів A чи L зі станів @. Таким чином ми отримуємо ПММ з вісьмома параметрами (чотири для переходу і чотири для результату), котра була параметризована, щоб мати властивості перераховані вище. Таку ПММ можна використовувати для видобування нелінгвістичних послідовностей для перевірки СКВ (стохастична контекстно-вільна граматика) моделлю. Потрібно звернути увагу на те, що такий підхід має обмеження, яке полягає в тому, що будуть обрізатися певні лінгвістичні символи котрі будуть розміщуватись на периферії емотикону.

Розроблювана система виконує індукцію СКВ граматики окремо в кожну послідовність з можливими емотиконами, базуючись на простому наборі методів шаблонів правил для присвоєння значимості правил. Індукуючи невеликі, окремі для кожного прикладу СКВ граматики, переконуємося, що кожний приклад має правильній розбір без збільшення граматики до розмірів котрі будуть впливати на ефективність аналізатора.

Сформована СКВ буде містити не термінальний X, та змінні a та b будуть вибрані з вхідної послідовності. Після вибору першої змінною вхідних даних є можливість вибору елементу котрий репрезентує середній сегмент. Тобто, можна зазначити, що СКВ містить два правила, виконання першого з них є необов'язковим. Отже, визначення елементу терміналу для участі в СКВ відбувається тільки збігом його з великим набором терміналів. Це дозволяє виконувати формування емотиконів, як от таких, що складаються тільки з двох не терміналів, а також таких, котрі є одним нетерміналом.

Описана базова індукція граматки може бути покращена кількома способами не жертвуючи надійністю цього методу . Один з них це розбиття вхідних символів на окремі слова. Другий – збільшення кількості символів нетерміналів в граматиці.