Т.И. Маджидов: Вы писали что в качестве "дескрипторов" строки использовались строки-токены, записанные как есть, строчными буквами, без первых символов. Какой эмбеддинг строк использовался? Это был one-hot или вы использовали какие-то языковые модели? Если one-hot, то почему? Если языковые модели, то какие? Каков был размер словаря?
Н.Ю. Бизюкова: На вход процедуры обучения/тестирования в используемой нами реализации (Python crfsuite) подаётся список токенов, конверсия их во внутреннее представление features, применяемых для обучения/распознавания терминов, производится встроенным алгоритмом процедуры CRF.

Т.И. Маджидов: Не очень понятна причина выбора отдельных "дескрипторов". Например, первый символ в строке.
Н.Ю. Бизюкова: Определение первого символа в строке было направлено на распознавание, в основном, наименований, которые являются систематическими названиями химических соединений. Часто первым символом в таких названиях является цифра. Это повышает точность распознавание метки “B”, если наименование химического соединения состоит из двух и более слов.
Сочетание указанных в презентации дескрипторов позволяет достигнуть максимальной точности прогнозирования из возможных при применении других дескрипторов.

Т.И. Маджидов: Как проводилась токенизация? Только по пробелам? Или по знакам препинания включительно? Если да, то как решалась проблема что название химического соединения может содержать скобки, знаки препинания и пр.?
Н.Ю. Бизюкова: Токенизация проводилась с помощью библиотеки nltk Python. Разделителями слов являются пробелы и знаки препинания (за исключением запятой без последующего пробела). При использовании этой библиотеки все знаки, которые послужили разделителями, кроме пробелов, сохраняются в списке, как токены, наравне с остальными словами. Таким образом, их влияние на распознавание химических соединений, содержащих знаки, было учтено.

Т.И. Маджидов: Сравнивались ли полученные модели с широко используемыми в настоящее время для NER, типа BERT?
Н.Ю. Бизюкова: Спасибо за предложение. Мы в настоящее время исследуем точность других методов, в том числе основанных на нейронных сетях, и приведём сравнение с BERT в будущем.