УДК 004.8 + 004.032.26

РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ С ИСПОЛЬЗОВАНИЕМ КОНТЕКСТНЫХ ВЕКТОРОВ

С.В. Слипченко

Международный научно-учебный центр информационных технологий и систем НАН Украины и Министерства образования и науки Украины

Исследовано распознавание именованных сущностей в CRF моделях по локальным контекстам без использования специализированных признаков. По результатам исследования для решения проблемы редких локальных контекстов предложен новый метод использования глобальной контекстной информации на основе распределенных представлений. Помимо интеграции глобальной контекстной информации, предложенный метод обеспечивает сокращение размерности и тем самым повышение эффективности обработки.

Досліджено розпізнавання іменованих сутностей у CRF моделях за локальними контекстами без використання спеціалізованих ознак. За результатами дослідження для вирішення проблеми мало поширених локальних контекстів запропоновано новий метод використання глобальної контекстної інформації на основі розподілених представлень. Крім інтеграції глобальної контекстної інформації метод забезпечує скорочення розмірності і таким чином збільшення ефективності обробки.

Введение

Постановка задачи. Одной из основных задач анализа текстов на естественном языке является выделение *именованных сущностей* [1–2] — имен людей, организаций, географических объектов и т.п. Например, в предложении «**[ORG** U.N.**]** official **[PER** Ekeus**]** heads for **[LOC** Baghdad**]**» содержатся три именованных сущности: организация "U.N.", человек "Ekeus" и географический объект "Baghdad". Выделенные сущности применяются как при последующем синтаксическом анализе и формировании семантических представлений, так и самостоятельно. Например, имена людей, организаций, географических объектов и т.п. индексируются и могут использоваться для поиска и фильтрации документов.

Перспективным направлением в развитии методов выделения именованных сущностей является использование априорной информации и нелокальных связей [3–4] в методах на основе максимизации энтропии [5] и условных случайных полей [3]. В большинстве моделей информативными признаками при распознавании именованных сущностей являются слова, которые встречаются в тексте в окрестности целевого (распознаваемого) слова, и их комбинации – локальные контексты. Проблемой распознавания именованных сущностей по локальным контекстам является то, что некоторые слова локальных контекстов могут очень редко встречаться совместно с распознаваемым словом. Это вызывает проблемы с обобщением при обучении моделей распознавания на ограниченной выборке.

Цели данной работы:

— исследовать качество распознавания именованных сущностей по признакам локальных контекстов и без традиционно используемых специализированных признаков;

— разработать новый подход к решению проблемы редких локальных контекстов путем использования глобальных контекстов, полученных на большой коллекции текстов [6].

Для сокращения размерности контекстных векторов предлагается использовать распределенное представления информации [7–10]. Это форма векторного представления информации, основанная на принципах нейросетевого представления информации в мозге, которая является реализацией подхода Николая Михайловича Амосова к моделированию мышления [11–12].

Методы решения задачи распознавания именованных сущностей

Задача выделения именованных сущностей в CoNLL-2003 [1] состоит в назначении каждому фрагменту теста одной из меток: **ORG**, **PER**, **LOC**, **MISC**, или **OTHER**, которую в простейшем случае можно рассматривать как назначение каждой лексеме (слову или знаку препинания) одной из перечисленных меток. Для обучения и тестирования алгоритма предоставляются предварительно размеченные обучающая и тестовая выборки.

Исторически методы распознавания именованных сущностей развивались от специализированных лингвистических методов для английского языка [2] до статистических методов с использованием обобщенных характеристик слов [1] (стиль написания, наличие спецсимволов, часть речи и т.п.). Впрочем, повсеместное использование смешанного регистра символов для распознавания именованных сущностей не позволяет использовать результаты статистических методов с локальными признаками в приложениях, где такая информация отсутствует. Например, в e-mail- или twitter-сообщениях не часто встречаются имена компаний, оформленные по всем правилам правописания.

Широко используемыми современными методами распознавания именованных сущностей являются методы на основе максимизации энтропии (MaxEnt) [5] и условных случайных полей (Conditional Random Fields, CRF) [3]. Оба этих метода используют экспоненциальную модель вероятности меток.

MaxEnt оценивает вероятность каждой метки независимо $P(y_i \mid \mathbf{x}_i, \mathbf{w}) = \frac{1}{Z(\mathbf{x})} e^{\mathbf{w}\cdot\mathbf{g}(y_i,\mathbf{x})}$, где $\mathbf{w}\cdot\mathbf{g}(y_i,\mathbf{x})$ — скалярное произведение вектора весов на вектор признаков, зависящих от последовательности слов $\mathbf{x}$ и метки текущего слова y_i. Наиболее вероятная последовательность меток определяется как последовательность наиболее вероятных меток в конкретной позиции, без учета меток в других позициях:

$$y_1^*,\ldots,y_n^* = \arg\max_{y_1 \in Y} P(y_1 \mid \mathbf{x},\mathbf{w}),\ldots,\arg\max_{y_n \in Y} P(y_n \mid \mathbf{x},\mathbf{w}).$$

CRF оценивает вероятность последовательности меток

$$P(y_i \mid y_{i-1},\mathbf{x},\mathbf{w}) = \frac{1}{Z(\mathbf{x})} e^{\mathbf{w}\cdot\mathbf{f}(y_{i-1},\mathbf{x},y_i)},$$

где $\mathbf{w}\cdot\mathbf{f}(y_{i-1},\mathbf{x},y_i)$ — скалярное произведение вектора весов $\mathbf{w}$ на вектор признаков $\mathbf{f}(y_{i-1},\mathbf{x},y_i)$, зависящих от метки предыдущего слова y_{i-1}, последовательности слов $\mathbf{x}$ и метки текущего слова y_i. Наилучшую последовательность для пар меток CRF определяет, используя динамическое программирование (детальнее см. Linear-Chain Conditional Random Fields в [13]):

$$y_1^*,\ldots,y_n^* = \arg\max_{y_1,\ldots,y_n \in Y\times\ldots\times Y} \prod_{i=1,\,n} P(y_i \mid y_{i-1},\mathbf{x},\mathbf{w}).$$

Таблица 1

Признаки CRF-моделей

Наименование	Описание
Базовые признаки	
w / wl	Текущее слово / текущее слово в нижнем регистре
shape	Сигнатура — текущее слово, в котором каждый символ заменен на соответствующую категорию (L — буква в нижнем регистре, U — буква в верхнем регистре, D — число и т.п.)
shaped	Аналогично **shape**, но последовательности категорий символов заменены на одну категорию
type	Тип слова (AllDigit, AllSymbol, AllAlnum и т.п.)
p1 – p4 / s1 – s4	Одно-, двух-, трех- и четырех буквенный префикс/суффикс слова
2d, 4d	Признак того, что слово является числом
d&a,d&-, d&/, d&,d&.	Различные комбинации цифр и знаков
up / iu	Буква в верхнем регистре, за которой следует точка / первая буква в верхнем регистре
au, al, ad, ao / cu,cl,ca,cs	Все символы в верхнем/нижнем регистре, цифры, не цифры или буквы / слово содержит букву в верхнем/нижнем регистре, букву или символ
Признаки учитывающие сходство слов в глобальном контексте	
wc	Класс/кластер, которому принадлежит слово в лексиконе

В большинстве случаев CRF [14] дает лучшие результаты, однако, как показывает ряд исследований [4], большую роль в качестве распознавания играет выбор признаков, используемых в модели. Обычно используются независимые наборы функций f — зависящие от метки в текущей позиции и последовательности символов $f' = \mathbf{1}_{y_i=\hat{y}_i}\phi'(i,\mathbf{x})$, а также меток в предыдущей и текущей позиции $f' = \mathbf{1}_{y_{i-1}=\hat{y}_{i-1},y_i=\hat{y}_i}\phi''(i,\mathbf{x})$, где $\mathbf{1}_\Diamond$ — индикаторная функция, принимающая значение 1 при выполнении условий $y_i = \hat{y}_i$ и

$y_{i-1} = \hat{y}_{i-1}, y_i = \hat{y}_i$ соответственно.

В табл. 1 приведены наиболее распространенные типы признаков φ, которые используются в CRF-моделях [15–16].

Таблица 2

Результаты CrfSuite и Stanford NER на CoNLL-2003

	CrfSuite Baseline Model			**Stanford NER (4 class distsim)**		
Type	**Recall**	**Precision**	**F1**	**Recall**	**Precision**	**F1**
ORG	0,8715	0,8528	0,8620	0,8770	0,8957	0,8862
MISC	0,8933	0,8347	0,8630	0,8662	0,9149	0,8899
PER	0,9355	0,9349	0,9352	0,9609	0,9470	0,9539
LOC	0,9126	0,9121	0,9123	0,9608	0,9525	0,9566
Avg	0,9032	0,8836	0,8931	0,9162	0,9275	0,92165

В табл. 2 представлены результаты CoNLL-2003 для CRF-моделей [15–16]. Распознавался стандартный набор классов именованных сущностей ORG — организации, MISC — другие, PER — люди, LOC — месторасположения. В качестве мер качества использовались стандартные меры Recall — полнота, Precision — точность, F1 — интегральная мера. Результаты **CrfSuite Baseline Model** (левый столбец) получены на базовом наборе признаков (см. табл. 1), а результаты **Stanford NER (4 class distsim)** (правый столбец) получены на расширенном наборе признаков, использующем класс/кластер слов лексикона. Расширенный набор признаков дает существенное улучшение качества.

КОНТЕКСТНЫЕ ВЕКТОРЫ И РАСПРЕДЕЛЕННЫЕ ПРЕДСТАВЛЕНИЯ

В современных методах представления и обработки векторной информации широко используются контекстные векторы. Элементами контекстного вектора слова являются значения некоторой функции от частоты совместной встречаемости этого слова с контекстом в некотором представительном корпусе текстов. В качестве контекстов могут использоваться различные компоненты текстов, например слова в окне определенной ширины вокруг слова или в одном предложении, абзаце, тексте и т.п.

Для формирования контекстных векторов слов требуется построить матрицу частот совместной встречаемости слов словаря и контекстов. При большом числе контекстов размерность этой матрицы может быть велика и она может не помещаться в оперативную память. Для сокращения размерности контекстных векторов можно использовать распределенные представления.

Распределенное представление (РП) информации — форма векторного представления, где каждый объект представлен множеством компонентов вектора и каждый компонент вектора может принадлежать представлениям многих объектов. Проекционные методы формирования РП уменьшают

размерность исходных векторов путем их умножения на случайную матрицу (матрицу со случайно сгенерированными элементами). Для ряда типов случайных матриц показано, что выходные векторы сохраняют характеристики входных, такие как сходство (скалярное произведение, угол) или расстояние. Для непосредственного формирования матрицы РКВ (сокращенной размерности), можно использовать прием, известный как случайное индексирование (Random Indexing or Random Labels [6]).

Результаты

Распознавания именованных сущностей с использованием локального контекста. Для проверки качества распознавания именованных сущностей без использования регистра символов был проведен ряд экспериментов с использованием в качестве признаков локального контекста слов в нижнем регистре (признак **wl** в табл.1).

Результаты представлены в табл. 3. Обозначения следующие: **wl**[i-2, i+2] — слова в окне [-2, +2] вокруг текущего слова **wl**[i], а **wl**[i-2, i+2] + (**wl**[i-2, i+2]**, wl**[i]) — слова в окне [-2, +2] и их комбинации с текущим словом — (**wl**[i-2, i+2]**, wl**[i]). Сравнение с результатами табл. 2 показывает, что, несмотря на значительное падение полноты распознавания *recall*, сохраняется высокая точность распознавания *precision* (левая часть табл. 3). Использование комбинаций слов дает *precision* больше, чем в *CrfSuite Baseline* и Stanford *NER*, что представляет интерес для ряда приложений. Кроме того, скорость обработки при этом возрастает в несколько раз по сравнению с экспериментами из табл. 2.

Таблица 3

Результаты экспериментов с CrfSuite без использования регистра символов

Type	**w**[i-2, i+2]			**w**[i-2, i+2] ∪ (**w**[i-2, i+2]**, w**[i])		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
ORG	0,5827	0,8976	0,7067	0,6176	0,9163	0,9737
MISC	0,5599	0,9173	0,6954	0,6144	0,9523	0,7379
PER	0,7139	0,9301	0,8078	0,7183	0,9359	0,7469
LOC	0,7225	0,9421	0,8178	0,7755	0,9519	0,8128
Avg	0,6448	0,9218	0,7569	0,6815	0,9391	0,8178

Как упоминалось выше, анализ результатов табл. 2 показывает, что использование расширения набора слов-признаков локального контекста путем добавления сходных слов позволяет добиться улучшения качества распознавания. Такое расширение признаков может рассматриваться как использование более глобального контекста.

Использование контекстных векторов в качестве признаков глобального контекста. Формирование контекстных векторов слов можно рассматривать как суммирование признаков всех локальных контекстов, в которых они встречались в корпусе текстов, который использовался для формирования контекстных векторов. Таким образом, контекстные векторы

можно считать представлениями глобального контекста признака-слова.

Для слова в текущей позиции вычисление $\mathbf{w} \cdot \mathbf{f}(y_{i-1}, \mathbf{x}, y_i)$ можно представить следующим образом:

$$\mathbf{wf}(y_{i-1}, \mathbf{x}, y_i) = \sum_{\hat{y}_i, j} w_{\hat{y}_i, j} \mathbf{1}_{y_i = \hat{y}_i} \mathbf{1}_{x_i = X_j} + \sum_{\hat{y}_i - 1, \hat{y}_i, j} w_{\hat{y}_{i-1}, \hat{y}_i, j} \mathbf{1}_{y_{i-1} = \hat{y}_{i-1}, y_i = \hat{y}_i} \mathbf{1}_{x_i = X_j} =$$

$$= \sum_{\hat{y}_i} \mathbf{1}_{y_i = \hat{y}_i} \sum_{\hat{y}_i, j} w_{\hat{y}_i, j} \mathbf{1}_{x_i = X_j} + \sum_{\hat{y}_{i-1}, \hat{y}_i} \mathbf{1}_{\hat{y}_{i-1}, \hat{y}_i} \sum_{\hat{y}_i, j} w_{\hat{y}_{i-1} \hat{y}_i, j} \mathbf{1}_{x_i = X_j} =$$

$$= \sum_{\hat{y}_i} \mathbf{1}_{y_i = \hat{y}_i} \mathbf{w}_{\hat{y}_i} \mathbf{X}_i + \sum_{\hat{y}_{i-1}, \hat{y}_i} \mathbf{1}_{\hat{y}_{i-1}, \hat{y}_i} \mathbf{w}_{\hat{y}_{i-1}, \hat{y}_i} \mathbf{X}_i,$$

где x_i — текущее слово, X_j — слово из лексикона, $\mathbf{X}_i$ — вектор, соответствующий текущему слову ($\mathbf{X}_{i,j} = 1$ если $i = j$, иначе 0), а $\mathbf{w}_{\hat{y}_i}$ и $\mathbf{w}_{\hat{y}_{i-1}, \hat{y}_i}$ — фрагменты вектора $\mathbf{w}$, соответствующие весам для меток $\hat{y}_{i-1}$ и $\hat{y}_i$.

Для представления глобального контекста заменим векторы слов $\mathbf{X}_i$ на контекстные векторы слов $\widetilde{\mathbf{X}}_i$, в результате получим:

$$\mathbf{wf}(y_{i-1}, \mathbf{x}, y_i) = \sum_{\hat{y}_i} \mathbf{1}_{y_i = \hat{y}_i} \widetilde{\mathbf{w}}_{\hat{y}_i} \widetilde{\mathbf{X}}_i + \sum_{\hat{y}_{i-1}, \hat{y}_i} \mathbf{1}_{\hat{y}_{i-1}, \hat{y}_i} \widetilde{\mathbf{w}}_{\hat{y}_{i-1}, \hat{y}_i} \widetilde{\mathbf{X}}_i .$$

Представления слов в позициях $-L, \ldots, 0, \ldots, +L$ в простейшем случае можно получить конкатенацией соответствующих контекстных векторов слов $\widetilde{\mathbf{X}}_{-L}, \ldots, \widetilde{\mathbf{X}}_0, \ldots, \widetilde{\mathbf{X}}_{+L}$.

Размерность конкатенированных КВ можно сократить, используя описанный выше метод случайных проекций, и получим распределенные КВ, представляющие глобальный контекст, которые будем сокращенно называть РКВГ.

Полученные случайным проекционным преобразованием РКВ являются неразреженными и небинарными. Необходимо исследовать работу разных реализаций CRF для подобных случаев. Также представляет интерес применение бинарных РКВ — возможно, с соответствующей модификацией вычисления $\mathbf{wf}(y_{i-1}, \mathbf{x}, y_i)$.

Выводы

Применение CRF-моделей для распознавания именованных сущностей в текстах при использовании слов локального контекста в качестве признаков дает достаточно высокую полноту распознавания (recall) и очень высокую точность (precision). Для дальнейшего улучшения результатов распознавания предложен подход, сочетающий глобализацию контекста путем использования контекстных векторов, а также преобразование векторных представлений признаков в распределенные представления. Направлением дальнейших исследований является реализация и экспериментальная проверка различных вариантов предложенного подхода.

1. Tjong E.F., Sang K., De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proc. of CoNLL-2003*. Edmonton, 2003, Canada, pp. 142–147.
2. Grishma R. Design of the MUC-6 Evaluation. *Proc. of the 6th Message Understanding Conf.* Maryland, Columbia, 1995, pp. 1–11.
3. Finkel J. R., Grenager T., Manning C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proc. of the 43nd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, Michigan, 2005. pp. 363–370.
4. Ratinov L., Roth. D. Design Challenges and Misconceptions in Named Entity Recognition. *Proc. of the 13th on Computational Natural Language Learning (CoNLL 2009)*. Boulder, Colorado, 2009, pp. 147–155.
5. Curran J. R., Clark. S. Language Independent NER using a Maximum Entropy Tagger. *Proc. of CoNLL-2003*. Edmonton, 2003. Canada, pp. 164–167.
6. Мисуно И.С. Векторные и распределенные представления, отражающие меру семантической связи слов / И.С. Мисуно, Д.А. Рачковский, С.В. Слипченко // Математические машины и системы. — 2005. — № 3. — С. 50–67.
   Misuno I.S., Rachkovskij D.A., Slipchenko S.V. Vector and distributed representations reflecting semantic relatedness of words. *Mathematical Machines and Systems*, 2005, Issue 3, pp. 50–67.
7. Куссуль Э.М. Ассоциативные нейроподобные структуры / Э.М. Куссуль. — К. : Наукова думка, 1992. — 144 с.
   Kussul E.M. *Associative neuron-like structures*. Kiev: Naukova Dumka, 1992. 144 p.
8. Rachkovskij D.A., Kussul E. Binding and normalization of binary sparse distributed representations by context-dependent thinning. *Neural Computation*, 2001, vol. 13, no. 2, pp. 411–452.
9. Rachkovskij D.A., Kussul E.M., Baidyk T.N. Building a world model with structure-sensitive sparse binary distributed representations. *Biologically Inspired Cognitive Architectures*, 2013, vol. 3, pp. 64–86.
10. Рачковский Д.А. Рандомизированные проекционные методы формирования бинарных разреженных векторных представлений / Д.А. Рачковский, И.С. Мисуно, С.В. Слипченко // Кибернетика и системный анализ. — 2012. — № 1. — С. 176–188.
    Rachkovskij D.A. , Misuno I.S., Slipchenko S.V. Randomized projective methods for the construction of binary sparse vector representations. *Cybernetics and Systems Analysis*, 2012, vol. 48, no. 1, pp. 146–156.
11. Амосов Н.М. Моделирование мышления и психики / Н.М. Амосов. — К.: Наукова думка, 1965, 304 с.
    Amosov N.M. *Modelling of thinking and the mind*. New York: Spartan Books, 1967. 304 p.
12. Нейрокомпьютеры и интеллектуальные роботы / Н.М. Амосов, Т.Н. Байдык, А.Д. Гольцев и др. — К. : Наукова думка, 1991. — 269 с.
    Amosov N.M., Baidyk T.N., Goltsev A.D., Kasatkin A.M., Kasatkina L.M., Kussul E.M., Rachkovskij D.A. *Neurocomputers and intelligent robots*. Kiev: Naukova dumka, 1991. 269 P.
13. Sutton C., Mccallum A. An Introduction to Conditional Random Fields for Relational Learning. *Introduction to Statistical Relational Learning*. Edited by L. Getoor, B. Taskar. Cambridge: MIT Press, 2007, pp. 93–128.
14. Agarwal M. , Goutam R., Jain A., Kesidi S.R. Comparative Analysis of the Performance of CRF, HMM and MaxEnt for Part-of-Speech Tagging, Chunking and Named Entity Recognition for a Morphologically rich language. *Proc. of Pacific Association For Computational Lingustics*. Kuala Lumpur, Malaysia, 2011, pp. 3–6.
15. Okazaki N. *CRFsuite - A fast implementation of Conditional Random Fields (CRFs)*. Available at: http://www.chokkan.org/software/crfsuite (Accessed 1 April 2013).
16. Finkel J., Klein D., Manning C. *Stanford Named Entity Recognizer (NER)*. Available at: http://nlp.stanford.edu/software/CRF-NER.shtml (Accessed 1 April 2013).

Получено 15.07.2013