РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ В ТЕКСТАХ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ С ИСПОЛЬЗОВАНИЕМ МЕТОДА ПРОБРОСО-ЦЕПОЧНЫХ УСЛОВНЫХ СЛУЧАЙНЫХ ПОЛЕЙ

Лабутин Иван Александрович (barracuda72@bk.ru)
Фирсов Антон Николаевич (a_firsov@mail.ru)
Чуприна Светлана Игоревна (chuprinas@inbox.ru)

Кафедра математического обеспечения вычислительных систем

Пермский государственный национальный исследовательский университет, 614990, Россия, г. Пермь, ул. Букирева, 15

This paper presents application of skip-chain conditional random fields to named entity recognition task. Skip-chain CRF, unlike linear approaches (such as LC-CRF and HMM) is able to represent long-distance relationships between text objects. Existing methods of capturing such connections has certain shortcomings. This work is dedicated to new method based on computing feature vectors difference norm. It removes unneccesary payload of data processing and makes the algorithm independent from specific natural language. The proposed method is implemented in GRMM package from MALLET and tested on freely available data.

Keywords: information extraction, named entity, IE, NER, CRF

Введение

Извлечение информации - группа задач, позволяющих по тексту на естественном языке в свободном формате получить некоторое структурированное представление информации, в нем содержащейся [1]. Задача выделения именованных сущностей является одной из ключевых, позволяя узнать, о каких именно объектах идет речь в тексте.

Традиционно в задачах извлечения сущностей применяются подходы, основанные на технологиях машинного обучения [2]. Одними из первых были скрытые марковские модели (HMM) [3]. К сожалению, HMM имеют один серьезный недостаток - они относятся к генеративным моделям, то есть к классу методов, основывающихся на моделировании функции совместного распределения $p(y, x)$, что при практическом применении приводит к трудностям при попытке отражения большого количества сложных взаимосвязей во входных данных: проблема вывода для них в общем случае неразрешима [4]. Кроме того, они, как правило, накладывают жесткое ограничение, требующее независимости факторов, описывающих взаимосвязи во входных данных.

Эти проблемы решаются в дискриминативных методах, которые моделируют функцию условного распределения $p(y|x)$. Основным плюсом данного подхода [5] является то, что условное распределение $p(y|x)$ не включает в себя модель для $p(x)$, которая и не требуется для задач классификации. К популярным методам данного класса относятся марковские модели максимальной энтропии (MEMM). Подходы этой группы демонстрируют более высокую производительность по сравнению с генеративными моделями, но им присущ другой недостаток - проблема “высокого порога” (high bias)[4].

Условные случайные поля (Conditional Random Field, CRF) [4] - класс статистических методов, использующих в качестве вероятностной модели условного распределения ненаправленный граф зависимости значения исследуемой случайной величины Y от наблюдаемой X. В отличие от методов группы MEMM, моделированию подвергается функция совместного распределения состояний, а не частные функции распределения каждого элемента из этого множества. За счет подобной модификации переходы между состояниями соревнуются в рамках

всей модели, что принципиально решает проблему “высокого порога”. Это делает CRF одним из самых практически интересных на сегодняшний день подходов [5].

В рамках CRF выделяются несколько различных вариаций, и наиболее популярными в задачах извлечения информации является линейно-цепочные условные случайные поля (LC-CRF) [5]. Они не уступают по выразительности методам HMM и MEMM, а по вычислительной эффективности - превосходят их [4]. Но бóльший практический интерес именно в задачах NER представляют пробросо-цепочные условные случайные поля (SC-CRF) [5], которые позволяют отражать “удаленные” связи во входных данных, такие как, например, разные вхождения одной и той же сущности в рамках новостной статьи. Однако, здесь встает новая проблема - протягивание этих самых “удаленных” связей, или, более точно, определение пар объектов, между которыми эти связи необходимо протягивать.

Существующие варианты построения “удаленных” связей в SC-CRF (наивный[5], с использованием регулярных выражений[6], словарный[7], классификатор [8]) имеют те или иные недостатки: языкозависимость[5], необходимость ручной обработки данных и/или программирования дополнительной логики[6], привлечение квалифицированных специалистов в области лингвистики[5][6].

Данная статья представляет решение проблем, присущих вышеперечисленным методам - связывание состояний модели на основе нормы разности векторов факторов [7]. Алгоритм переиспользует векторы факторов элементов текста, полученные на более ранних этапах конвейера извлечения информации, при этом оперируя ими в обезличенном виде, как с обычными математичскими векторами. Это обеспечивает независимость от языка входного текста и отсутствие необходимости дополнительной ручной работы (подготовки словарей и составления реуглярных выражений), что повышает удобство использования SC-CRF и потенциально может улучшить качество извлечения сущностей.

1. Условные случайные поля

Формальное определение CRF приведено в [4]:
Пусть $G = (V, E)$ - граф, такой, что каждая его вершина v соответствует некоторому состоянию $Y_v \in Y$. Тогда (X, Y) - условное случайное поле, если случайные величины Y_v, подчиняясь условному распределению по X, удовлетворяют марковскому свойству по отношению к графу:

$$p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$$

Другими словами, вероятность оказаться в любой вершине Y_v графа никак не зависит от всех остальных вершин из множества Y. Таким образом, CRF есть случайное поле, полностью определяемое множеством значений наблюдаемой величины X.

В случае, когда граф $G = (V, E)$ является достаточно простым (деревом или цепочкой), авторы [4] определяют вероятность появления определенной последовательности меток y с учетом имеющихся наблюдений x как нормированное произведение потенциальных функций, каждая из которых имеет вид:

$$f_j(y_{i-1}, y_i, x, i) = exp(\sum_{e \in E, j} \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_{v \in V, k} \mu_k s_k(y_i, x, i))$$

Здесь $t_j(y_{i-1}, y_i, x, i)$ - функция факторов перехода, зависящая от всех наблюдений и меток на текущей и предыдущей позициях; $s_k(y_i, x, i)$ - функция факторов состояний, зависящая от всех наблюдений и метки на текущей позиции; λ_j и μ_k - параметры, получаемые путем обучения алгоритма.

При определении факторов, участвующих в работе алгоритма, составляются веществннозначные функции $b(x, i)$, которые отражают некоторые характеристики эмпирического распределения. Каждая функция факторов принимает значение одной из данных веществннозначн-

ных функций, если текущее (и предыдущее, в случае функции переходов) состояния достигается на некоторых значениях входных данных. Например,

$$t_j(y_{i-1}, y_i, x, i) = \begin{cases} b(x,i) & \textit{если } y_{i-1} = IN \textit{ и } y_i = INP \\ 0 & \textit{в противном случае} \end{cases}$$

Если обозначить фукнции факторов переходов t_j и функции факторов состояний s_k как f_j, то итоговую вероятность появления определенной последовательности меток y с учетом имеющихся наблюдений x можно записать следующим образом[9]:

$$p(y|x, \lambda) = \frac{1}{Z(x)} exp(\sum_j \lambda_j F_j(y, x))$$

Здесь $F_j(y, x) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x, i)$ (i здесь пробегает последовательность меток), $Z(x)$ - нормализующая фукнция.

В практических приложениях при поиске параметров λ_j, как правило, производится максимизация логарифма функции правдоподобия[9]:

$$L(x) = \sum_k \Big[\sum_j \lambda_j F_j(y^{(k)}, x^{(k)}) - log(Z(x^{(k)})) \Big] \quad (1)$$

(Здесь k пробегает множество обучающих примеров.) Данная функция является выпуклой, благодаря чему гарантирована сходимость к глобальному минимуму. Однако, аналитически эта задача в общем случае не решается. На практике используются итеративные методы: различные разновидности алгоритма Бройдена - Флетчера - Гольдфарба - Шанно (BFGS), например, BFGS с ограничением по памяти (L-BFGS [10]); вариации метода градиентного спуска, такие, как стохастический градиентный спуск (SGD [11]); усредненный перцептрон (AP [12]) и другие.

Выше была приведена модель CRF для простого случая, когда CRF является деревом или цепочкой (Linear-Chain CRF). Но на практике зачастую возникает потребность в использовании более сложных структур. Кроме того, CRF (в сравнении с HMM), как дискриминативная модель, показывает себя наиболее хорошо именно в ситуациях сложных зависимостей. Первым шагом в этом направлении являются пробросо-цепочные условные случайные поля (Skip-Chain CRF) [5].

Пробросо-цепочные CRF, или SC-CRF возникают как обобщение последовательно-цепочных (Linear-Chain CRF, LC-CRF) путем ухватывания не только связей между соседними объектами, но и более отдаленных отношений. Применительно к извлечению информации они представляют довольно простую идею: “похожие” слова/словосочетания в разных частях небольшого по объему текста (новостная статья, объявление), скорее всего, обозначают одну и ту же сущность. Например, в новости, посвященной подписанию договора между двумя странами, словосочетания “китайская сторона”, “Китай”, “правительство Китая” в схожих контекстах с большой долей вероятности будут обозначать одно и то же. В качестве преимущества данного подхода можно отметить тот факт, что каждое вхождение сущности в текст в одной из своих форм может нести дополнительную смысловую нагрузку, повышающую качество работы алгоритма за счет увеличения количества обрабатываемых факторов. Исследования на аглоязычных текстах [5][13] показали, что SC-CRF в целом осуществляет задачу извлечения сущностей лучше, чем LC-CRF; особенно хорошо этот алгоритм проявляет себя при идентификации информации о людях (имя, отчество) и географических объектах.

Однако за более выразительную модель приходится платить повышением вычислительной сложности задачи обучения. В графе SC-CRF могут присутствовать циклы, которые, к тому же, могут иметь общие части; это переводит задачу в класс NP-полных[6]. На практике вместо

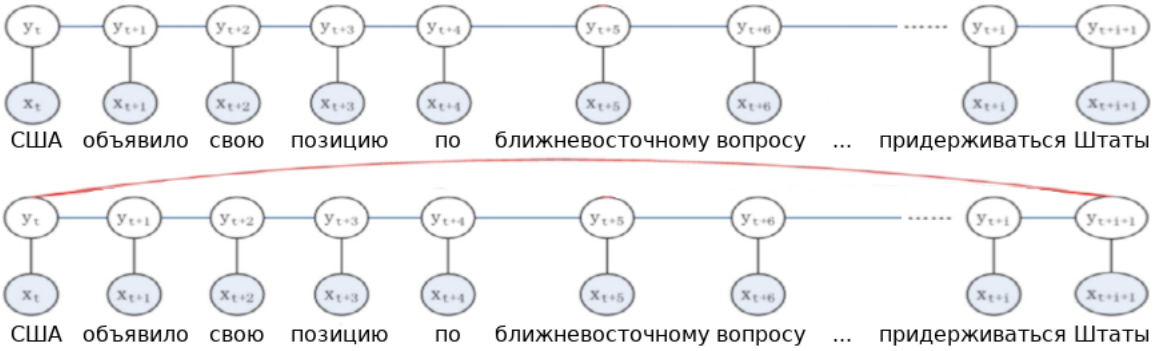

Рис. 1: Пример LC-CRF (сверху) и SC-CRF (снизу)

(1) используется приближенное правдоподобие Бете [14]:

$$\hat{L}(\Lambda; \{b\}) = \sum_k log\Big[\frac{\prod_t \prod_c b_{t,c}(y_{t,c}^{(k)})}{\prod_s b_s(y_s^{(k)})^{d_s-1}}\Big] \tag{2}$$

(здесь c пробегает все клики, t - все переходы, s - все состояния, а d_s показывает степень вершины s или, другими словами, количество зависящих от нее факторов) в совокупности с алгоритмом L-метода. К сожалению, здесь возникает другая проблема: тогда как лог-функция максимального правдоподобия (1) выпукла, ее приближение (2) уже таковым не является. Это означает, что при поиске численного решения можно вместо глобального экстремума оказаться в локальном. Более того, локальных максимумов может быть несколько, и наборы полученных при обучении параметров Λ, соответствующие каждому из максимумов, будут давать разные результаты при работе. В связи с этим встает вопрос о начальной инициализации параметров алгоритма. Хорошей практической рекомендацией [14] будет использование для SC-CRF в качестве стартовых значений параметров, полученных при обучении соответствующего ему LC-CRF алгоритма на тех же исходных данных; параметры, соответствующие отсутствующим в модели LC-CRF связям, можно положить равными нулю.

2. Метрика схожести

Как отмечалось в предыдущем разделе, SC-CRF отличаются от LC-CRF тем, что позволяют представлять “удаленные” зависимости между сходными объектами текста. Интерес представляет как раз определение “схожести”. Авторы оригинальной модели, хоть и предлагают различные метрики (по схожести написания, по идентичности стема), в своей работе ограничиваются “наивным” критерием: два слова схожи, если они пишутся одинаково и начинаются с заглавной буквы [5]. В других исследованиях предлагаются более сложные варианты, например, с использованием знаний о грамматической структуре предложения (скажем, постулирование схожести слов, связанных союзом “и”) или о частях речи [6]. Таким образом, на данный момент существуют следующие основные варианты [7]:

1. “Наивное” [5]: соединение одинаковых по написанию слов, начинающихся с заглавной буквы. Простой вариант, имеющий, впрочем недостатки - различные написания одного и того же объекта уже не рассматриваются как кандидаты на связывание.

2. С использованием регулярных выражений [6]: модификация предыдущего варианта, способная улавливать вариации в написании. Требует составления собственно регулярных выражений, то есть, ручной работы, и, как и предыдущий вариант, делает алгоритм языкозависимым.

3. Словарное [7]: связывание объектов с использованием словаря синонимов. Позволяет увязывать различные имена одного и того же объекта и, в некоторой степени, языконезависима, однако узкое место - словарь, требующий составления.

4. Обучение классификатора [8]: рассмотрение данной подзадачи как полноценной задачи классификации и обучение некоторого классификатора для определения принадлежности слов одной группе. Метод достаточно универсален, но требует значительных вычислительных ресурсов[8].

В данной работе предлагается иной вариант определения похожести объектов текста - норма разности векторов факторов состояний. Основная идея состоит в том, что два состояния связываются между собой, если вектора факторов их наблюдений достаточно близки (в смысле евклидовой или некоторой другой метрики):

$$I(x, y) = d(\overline{b_x}, \overline{b_y}) < T$$

где I - функция, определяющая необходимость связывания объектов x и y, $\overline{b_x}$, $\overline{b_y}$ - векторы факторов объектов, d - некоторая метрика, T - порог, до которого объекты считаются достаточно похожими для протягивания между ними связи. Подобный метод часто используется в других задачах классификации, однако применительно к установлению связей в SC-CRF можно считать его новшеством. Такая метрика теоретически позволяет улавливать слова, не являющиеся синонимами в полном смысле этого слова, но употребляемые в одном и том же качестве в рамках текста. Кроме того, этот метод не полагается ни на какие внешние артефакты, типа словарей и не добавляет дополнительных накладных расходов на этапе протягивания связей - к этому моменту значения векторов факторов b_x и b_y уже вычислены. Эти особенности дают данному методу конкурентное преимущество по сравнению с имеющимися решениями, обеспечивая в том числе независимость от языка входного текста.

3. Тестирование и результаты

С целью сравнения предлагаемого метода с другими существующими реализациями код пакета GRMM из состава фреймворка MALLET был расширен алгоритмом, ответственным за создание связей в графе переходов между “достаточно близкими” состояниями; в качестве меры близости выступает абстрактный метод, определяемый в наследуемых классах, реализующих конкретные метрики: “наивную”, синонимов и векторную (в качестве основы векторной метрики использовалась евклидова норма). Кроме того, при тестировании использовались стандартные возможности фреймворка, предоставляющие возможность построения LC-CRF.

В системе используются следующие свойства элементов входного текста: 1. Словарная форма слова 2. POS-тег слова (часть речи) 3. Признак начала или конца предложения 4. Написание: с большой буквы, прописными буквами, строчными буквами.

Первоначальная тренировка модели в соответствии с рекомендациями [14] производилась без использования дополнительных дуг, свойственных модели SC-CRF. В дальнейшем модель расширялась данными дугами и производилась повторная тренировка модели с использованием полученных ранее параметров.

Алгоритмом обучения для получения параметров модели послужил вариант BFGS с ограничением по памяти (L-BFGS). В качестве критерия сходимости использовалось приближенное правдоподобие Бете.

В качестве словаря синонимов использовался подготовленный для машинной обработки словарь “Словарь русских синонимов и сходных по смыслу выражений” Н. А. Переферковича [15][16].

При тестировании в текстах выделялись следующие сущности: • географические наименования (Geox); • названия организаций (Orgn); • фамилии (Surn), имена (Name) и отчества (Patr) людей.

Следует отметить, что данное тестирование для простоты и возможности переиспользования имеющихся инструментов в составе фреймворка MALLET ограничивалось сущностями, состоящими из одного слова, однако это не является принципиальным ограничением предложенного метода.

Для обучения алгоритма и тестирования его эффективности использовались уже готовые размеченные данные Открытого корпуса русского языка OpenCorpora.org [17], подкорпус со снятой омонимией ($\sim$ 2000 текстов, состав корпуса: 660 Geox, 60 Orgn, 400 Name, 230 Surn, 34 Patr).

Таблица 1: Результаты алгоритмов, F1-score

Алгоритм	Geox	Orgn	Name	Surn	Patr	Общий
CRF++ 0.58 (LC-CRF)	0.89	0.82	0.86	0.88	0.85	0.83
CRFSuite 0.12 (LC-CRF)	0.86	0.82	0.83	0.80	0.78	0.82
MALLET (LC-CRF)	0.87	0.79	0.81	0.85	0.80	0.81
MALLET (SC-CRF, Naive)	0.86	0.81	0.82	0.86	0.78	0.84
MALLET (SC-CRF, Naive+Synonim)	0.84	0.86	0.86	0.85	0.80	0.85
MALLET (SC-CRF, Naive+Vector)	0.89	0.87	0.92	0.86	0.82	0.86
MALLET (SC-CRF, Все три)	0.93	0.91	0.94	0.92	0.84	0.88

Приведенные результаты проверены парным зависимым тестом Стьюдента (за baseline взяты результаты CRF++), на основании чего они признаны статистически значимыми с уровнем значимости 5%.

На основе полученных результатов можно сделать вывод о том, что предложенный метод на основе векторной метрики работает не хуже метрики синонимов, позволяя получить лучшие результаты и не требуя при этом применения внешних словарей, большого объема ручной работы, необходимости коррекции программной реализации под особенности конкретного языка и привлечения квалифицированных специалистов в области лингвистики. Также можно видеть, что наилучшие результаты дает комбинация всех трех методов.

Заключение

Представленный в данной статье метод определения схожести объектов текста на основании нормы разности векторов факторов позволиляет избавиться от проблем, присущих другим методам соединения состояний в рамках модели SC-CRF - языкозависимости, большого объеме ручной работы, необходимости коррекции программной реализации под нужды конкретной задачи и привлечения квалифицированных специалистов в области лингвистики. Показано, что построенная с использованием данного метода система не уступает имеющимся реализациям в извлечении именованных сущностей, при этом не требуя повышенных вычислительных затрат и не прибегая к использованию вручную составленных артефактов (словарей и регулярных выражений).

Список литературы

[1] Cowie Jim, Lehnert Wendy. Information Extraction // Commun. ACM. — 1996. — Январь. — Т. 39, № 1. — С. 80–91. — URL: http://doi.acm.org/10.1145/234173.234209.

[2] Indurkhya Nitin, Damerau Fred J. Handbook of Natural Language Processing. — Second edition. — Taylor and Francis Group, LLC, 2010. — P. 21–131, 511–532.

[3] Baum Leonard E., Petrie Ted. Statistical Inference for Probabilistic Functions of Finite State Markov Chains // Ann. Math. Statist. — 1966. — 12. — Vol. 37, no. 6. — P. 1554–1563. — URL: http://dx.doi.org/10.1214/aoms/1177699147.

[4] Lafferty John, McCallum Andrew, Pereira Fernando C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data // Proceedings of the 18th International Conference on Machine Learning 2001. — 2001. — June. — P. 282–289.

[5] Sutton Charles, McCallum Andrew. An Introduction to Conditional Random Fields for Relational Learning // Introduction to Statistical Relational Learning. — MIT Press, 2006.

[6] Liu Jingchen, Huang Minlie, Zhu Xiaoyan. Recognizing Biomedical Named Entities using Skip-Chain Conditional Random Fields // Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010. — 2010.

[7] Лабутин И. А. Извлечение сущностей из текстов на естественном языке с использованием метода Conditional Random Fields : Курсовая работа / И. А. Лабутин. — 2016.

[8] Galley Michel. A Skip-Chain Conditional Random Field for Ranking Meeting Utterances by Importance // Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. — 2006.

[9] Wallach Hanna M. Conditional Random Fields: An Introduction // University of Pennsylvania CIS Technical Report MS-CIS-04-21. — 2004.

[10] Nocedal Jorge. Updating Quasi-Newton Matrices with Limited Storage // Mathematics of Computation. — 1980. — P. 773–782.

[11] Pegasos: Primal Estimated sub-GrAdient SOlver for SVM / Shai Shalev-Shwartz, Yoram Singer, , Nathan Srebro // Proceedings of the 24th International Conference on Machine Learning (ICML 2007). — 2007. — P. 807–814.

[12] Collins Michael. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002). — 2002. — P. 1–8.

[13] Finkel Jenny Rose, Grenager Trond, Manning Christopher. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling // Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL). — 2005.

[14] Sutton Charles. Efficient Training Methods for Conditional Random Fields : Ph. D. thesis / Charles Sutton ; Graduate School of the University of Massachusetts Amherst. — 2008. — P. 42–93.

[15] Переферкович Наум Абрамович. Словарь русских синонимов и сходных по смыслу выражений. — 1890. — URL: https://nlpub.ru/Словарь_Абрамова.

[16] Переферкович Н. А. Словарь русских синонимов и сходных по смыслу выражений. — 1999.

[17] Алексеева Светлана, Бодрова Анастасия, Бочаров Виктор и др. OpenCorpora: открытый корпус русского языка. — 2016. — URL: http://opencorpora.org.