УДК 681.3

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ ТЕКСТА: ПРОГНОЗИРОВАНИЕ СВЯЗЕЙ МЕЖДУ ЗАДАННОЙ ПАРОЙ СУЩНОСТЕЙ

Манучарян Л.А.

Воронежская государственная лесотехническая академия, Воронеж, Россия, e-mail: levonmanucharian@gmail.com

В данной статье приведены примеры методов прогнозирования связей между заданной парой сущностей. Если дано фиксированное R-множество типов связей, каждая из которых включает пару типов сущностей, нашей целью является определение всех соответствий связей в R в заданном тексте на естественном языке, в котором помечены все сущности. Задача прогнозирования более легкая по сравнению с задачей по извлечению сущностей, так как в данном случае необходимо сделать только скалярный прогноз вместо векторного. Рассматриваются поверхностные репрезентации, подчеркивается важная роль тегов части речи и приводится сравнительный анализ дерева грамматического разбора и графа зависимостей и сценарий использования последних для задач извлечения связей. Далее приводится краткий обзор популярных типов извлечения и предлагается методика извлечения сущностей.

Ключевые слова: связь между парой сущностей, извлечение информации.

EXTRACTION OF INFORMATION FROM TEXT: PREDICTING THE RELATIONSHIP BETWEEN GIVEN ENTITIES.

Manucharyan L.A.

Voronezh state academy of forestry and technologies, Voronezh, Russia, e-mail: levonmanucharian@gmail.com

In this article the methods of predicting the relationship between a given entity pair are provided. Given a fixed set R of relationship types each having a pair of entity types, the final aim is to identify all occurrences of the relationships in R in an input natural language text in which all entities have been marked. The task of the prediction is easier than the task of the extraction of entities, because the latter case involves only scalar prediction instead of the vector one. Surface tokens are reviewed, the important role of part of speech tags is accentuated and a compare analysis between dependency graphs and parse trees is provided for various scenarios of extraction tasks. A short review of popular types of extraction is provided further and an extraction method of entities is suggested.

Key words: Entity pair relationship, information extraction.

Введение

В основном при извлечении связей из текста на естественном языке подразумевается, что две аргументные сущности находятся в непосредственной близости или же являются частью одного и того же предложения. Таким образом, базовая задача распознавания имеет следующий вид: если на входе имеется фрагмент текста x и две помеченные сущности E_1 и E_2 в x, нужно определить, существует ли связь Y между E_1 и E_2. Множество Y указывает на все типы связей R, а также на специальный тип «другое» для случая, когда ни одна из связей не применима к паре сущностей. Данная задача распознавания более легкая по сравнению с задачей по извлечению сущностей, так как в данном случае необходимо сделать только скалярный прогноз вместо векторного. Однако извлечение связей считается более сложной проблемой, чем извлечение сущностей, так как нахождение связей между двумя словами в предложении требует применения искусной комбинации локальных и нелокальных подсказок с «шумом» из разнообразных синтаксических и семантических структур в предложении. Далее будет представлен анализ наиболее общих типов ресурсов, полезных для извлечения связей:

Анализ типов ресурсов для извлечения связей

Поверхностные репрезентации: репрезентации вокруг и между двумя сущностями часто включают в себя информацию, необходимую для извлечения связей. Например, связь «расположена» между сущностями «Компания» и «Расположение», четко обозначается присутствием репрезентаций N-граммы «расположена» и биграммы «расположена» между двумя сущностями, как в примере:

<Компания>Роснефть</ Компания > расположена по адресу

<Расположение>г. Москва, ул. Харькова</ Расположение >.

Аналогично, связь между «болезнью» и «расположением» четко выделяется присутствием слов вроде «эпидемия». Например – центр за контролем и предотвращением заболеваний выявил случаи эпидемии <Болезнь>чумы</Болезнь> в нескольких городах <Расположение>Египта</Расположение>.

Часто репрезентация обобщается или приводится к своему морфологическому корню. Напр., «расположен» приводится к слову «расположение».

Теги части речи. Теги части речи играют более существенную роль в извлечении связей, чем в извлечении сущностей. Глаголы в предложении являются ключевым фактором в определении связи между сущностями, которые в большинстве случаев представляют собой именные группы. Например в предложении – «В этом году < Расположение > Воронежская лесотехническся академия</ Расположение > выступит в роли принимающей стороны в проведении <Олимпиада>ПБГ </Олимпиада>» более достоверное извлечение связи «проведение» между «олимпиадой» и «расположением» было бы возможно, если бы фраза «принимающая сторона» была тегирована как глагол, вместо именной.

Структура дерева синтаксического разбора. Дерево синтаксического разбора группирует слова в предложении в выступающие типы грамматических оборотов, такие как именные группы, предложные группы и глагольные группы и таким образом представляет намного большую ценность в определении связей между сущностями в предложении, нежели теги POS (POS – Part Of Speech, Часть речи).

Например, в предложении «В <Расположение>Воронеже</Расположение>, находящемся в 500 километрах от <Расположение> Москвы </Расположение>, будет проводится олимпиада <Олимпиада>ПБГ</Олимпиада> в 2012 году» предпочтение, скорее всего, будет отдано «Москва, ПБГ», чем «Воронеж, ПБГ», как экземпляру связи «будет проводится», основанному на относительной близости к «ПБГ». Однако дерево грамматического разбора для выше иллюстрированного предложения приносит «ПБГ» ближе к слову «Воронеж», чем «Москва», так как «Воронеж» является главной именной в предложении «Воронеж, расположенный в 520 километрах от Москвы», который в свою очередь формирует подлежащее глагольной группы «будет проводится в 2010».

Граф зависимостей. Создание полноценного дерева разбора обходится дорого. Граф зависимостей, который связывает каждое слово со словами, находящимся в зависимости от него, зачастую может быть настолько полезным, насколько дерево разбора. Например, граф зависимостей для вышеописанного предложения показан на рис. 1. Из графа видно, что глагол «проводится» связан с обеими сущностями: «Воронеж» – сущность «Расположение» и «ПБГ» – сущность «Олимпиада». Это напрямую создает близкую связь между обеими сущностями.

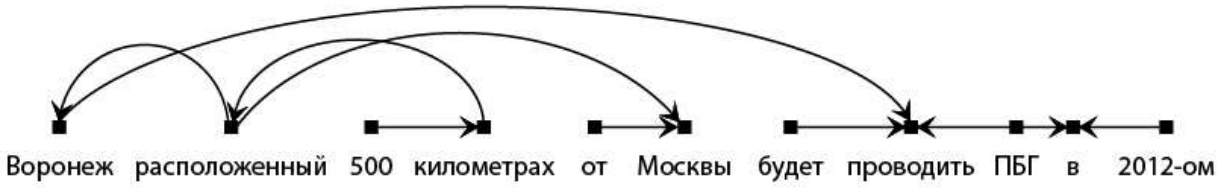

Рис. 1. Разбор зависимостей предложения.

В отличие от этого путь между «ПБГ» и «Москва» идет через «Воронеж» и «расположенный». Далее будут рассмотрены методы, которые используют вышеописанную информацию в разных вариациях для классификации входного (x, E1, E2) в одного из Y классов. Предположим, имеется N обучительных примеров в виде (x^i, E_1^i, E_2^i, r^i): i=1….N, где $r^i \in Y$ обозначает связь, которая существует между сущностями $E_1^{i,}$, $E_2^{i,}$ в предложении x^i. Главной проблемой является обработка разнообразных структурных видов, которые включают в себя разные входные данные. Например, репрезентации и теги части речи образуют последовательность, выходная информация при грамматическом разборе является деревом, а структура зависимостей – графом. Далее могли бы присутствовать ошибки в любом из входных ключей, так как лингвистические библиотеки, используемые для задач извлечения, несовершенны. Несмотря на то что избыточность придает устойчивость ошибкам, переизбыток может привести к информационному шуму и увеличить длительность работы приложения. Методы, используемые в извлечении связей, могут быть распределены по следующим трем главным типам.

- *Методы, ориентированные на топологию*: извлекается фиксированное множество дифференциальных признаков из входной информации с дальнейшим запуском готового классификатора (напр., дерево решений или SVM [4]).
- *Методы, базированные на ядре*: создаются специальные ядра для перехвата сходств между структурами вроде дерева или графов. Ядро в большинстве представляет собой ключевое слово в предложении.
- *Методы, основанные на правилах*: создаются пропозициональные правила и правила первого порядка для структур вокруг двух сущностей. Для примеров разных систем извлечения связей [1; 2].

Методика извлечения

Методы, ориентированные на топологию: большое разнообразие методов использовалось для конвертации ключей извлечения структур вида последовательностей, деревьев или графов, в

фиксированное множество признаков – для использования стандартными классификаторами. Авторы в [5] представляют систематический метод проектирования таких признаков, одновременно включая большинство ранее предложенных методов извлечения связей, основанных на признаках [6; 7].

Обозначим входное предложение X, где x_i обозначает слово в позиции i, а E_1, E_2 обозначают сегменты в x, соответствующие двум сущностям, связь которых желаем прогнозировать. Для простоты предположим, что и E_1 и E_2 состоят из одного слова. Каждое слово x_i связано со множеством свойств $p_1 \ldots p_k$, так же как признаки используются при извлечении сущностей X. Примеры таких свойств включают строчную форму x_i, орфографический тип x_i, класс x_i в заданной онтологии, помеченная сущность x_i, а также POS тег x_i. Первое множество функциональных возможностей приобретается путем перехвата всех возможных союзов свойств двух репрезентации, представляющих собой две сущности E_1 и E_2. Примеры таких случаев следующие.

[[Помеченная сущность от E_1 = Личность, Помеченная сущность от E_1 = Расположение]].

[[Строка E_1 = «Эйнштейн», Орфографический тип от E_2 =4-цифры]].

Первая функция могла бы быть полезной для связи «проживает в», когда первая сущность в предложении – «Личность», а вторая помеченая сущность – «Расположение».

Второй признак мог бы быть полезным для связи «рожден», когда E_1 равняется «Энштейн» и E_2 состоит из 4-х цифр. Далее представим, как нужно извлекать дифференциальные признаки из структурных входных данных, что включает связи между словами в предложении. Имеется три типа входных данных: последовательности, дерево разбора и графы зависимостей. Для объединения шага генерации дифференциальных признаков от этих входных данных каждый будет рассматриваться как граф, вершины которого являются словами, а также нетерминалы (NNP, VP, и т.д.) в случае с деревом грамматического разбора. Каждое слово-вершина ассоциировано с K множеством $p_1 \ldots . p_k$. Вдобавок специальный признак добавляется к каждой вершине, который может принять четыре значения: 1 – если относится к E_1, 2 – если относится к E_2, «обе» – если относится к обеим, и «никакой», если ни к которой не относится. Дифференциальные признаки являются производными от свойств индивидуальных вершин, а также пар вершин, связанных краем, или троицами вершин, связанных по крайней мере двумя краями, и для каждой комбинации значений спецпризнака, связанного с краями. Пример

вышеописанного может быть иллюстрирован при помощи предложения X с E_1= «Воронеж» и E_2= «ПБГ».

В <Расположение>Воронеже</Расположение>, расположенном в 520 километрах от <Расположение>Москвы</Расположение>, будет проведена олимпиада <Олимпиада>ПБГ</ Олимпиада>, которая состоится в 2012 году.

Признаки от последовательности слов: в первую очередь будут приведены примеры дифференциальных признаков для графа последовательности, преобразованного словами между E_1 и E_2. В этом случае единственные края находятся между смежными словами. Каждая вершина включает К признаков и один из возможных специальных признаков – (1, 2, «никакой»).

Примеры дифференциальных признаков n-граммы следующие.

[[строка =«проводить», спецпризнак =« никакой»]]

[[Часть речи = Глагол, спецпризнак =« никакой»]].

Примеры дифференциальных признаков биграмм следующие.

[[строка = «(проводить, ПБГ)», спецпризнаки= «(никакой, 2)», тип =«последовательность»]]

[[Часть речи = (Глагол, существительное), спецпризнаки= «(никакой, 2)», тип =«последовательность»]]

Примеры дифференциальных признаков триграмм следующие.

[[Строки = «(будет, проводить, «ПБГ»)», спецпризнаки= «(никакой, никакой, 2)», тип =«последовательность»]]

[[Часть речи = (модификатор, глагол, существительное), спецпризнаки= «(никакой, никакой, 2)», тип =«последовательность»]]

Используя этот шаблон, можно с легкостью рассчитать максимальное число дифференциальных признаков для каждого типа. Обозначим $d(p_i)$ число возмозных значений, которые может

принимать признак i, а также обозначим $d = \sum_{i=1}^{k} d(p_i)$ сумму величины всех признаков. Тогда число для N-грамм признаков будет 3d, биграмм – $3^2 d^2$, и для триграмм – $3^3 d^3$. На практике число признаков намного меньше, так как во время обучения учитывается только комбинация признаков, присутствующая хотя бы в одном экземпляре обучения.

Дифференциальные признаки от графа зависимостей. Рассмотрим граф зависимостей из рис. 1. Так как множество вершин то же, что и при последовательностях, больше не требуется генерировать n-грамм признаки. Края в графе создают множество биграмма и триграмма признаков. Вот несколько примеров.

[[(пометка сущности = «Расположение», Часть речи = (Глагол), спецпризнаки= «(1, никакой)», тип =«зависимость»]].

Этот признак запускается на паре вершин «(Воронеж, проводится)», связанных в графе зависимостей как «проводится ← Воронеж». Одно из полезных триграммных свойств, полученных из графа зависимостей, следующее:

[[(POS = (существительное, Глагол, существительное), спецпризнак = «(1, никакой, 2)», тип =«зависимость»]].

Это свойство запускается на вершинах «(Воронеж, проводится, «ПБГ»)», из-за шаблона края – «Воронеж → проводится ← ПБГ».

Дифференциальные признаки от дерева грамматического разбора. Множество вершин в дереве разбора состоит из вершин-слов в краях и внутренних вершин. Это открывает новые N-грамм свойства следующего вида:

[[вершина = «VP», спецпризнак = «2»]].

Также некоторые из внутренних вершин, которые относятся к сущностям вершин E_1 и E_2, теперь могут иметь спецпризнак «обе». Например, в рис. 1, вершина «S» относится к сущностям, а также ассоциирована со спецпризнаком «обе». Вдобавок, используя края от «родителя до детей» в дереве разбора, можно определить биграммные и триграммные признаки, как в случае с графами зависимостей. Авторы в [5] утверждают, что в восьми задачах по извлечению из ACE корпусов, достигнутая точность с этими простыми признаками, производными от N-грамм, Биграмм и триграмм от входных графов является

конкурентоспособной с другими методами, которые интерпретируют структуру более глобально.

Заключение

В данной статье были рассмотрены методы прогнозирования связей между заданной парой сущностей при помощи анализа существующих методов и экспериментальным путем были выявлены наиболее эффективные варианты извлечения.

Список литературы

1. Aitken J. Learning information extraction rules: An inductive logic programming approach // Proceedings of the 15th European Conference on Artificial Intelligence, 2002, p. 355–359.
2. Jayram T.S., Krishnamurthy R., Raghavan S., Vaithyanathan S. and Zhu H. Avatar information extraction system // IEEE Data Engineering Bulletin, 2006, V. 29, p. 40–48.
3. Marie-Catherine de Marnee and Christopher D. Manning. Stanford Typed Dependencies Manual. p. 2-11, sept. 2008.
4. Метод опорных векторов (SVM). – URL: http://ru.wikipedia.org/wiki/SVM.
5. Jiang J. and Zhai C. A systematic exploration of the feature space for relation extraction // Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, 2007, p. 113–120.
6. Kambhatla N. Combining lexical, syntactic and semantic features with maximum entropy models for information extraction // The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain: Association for Computational Linguistics, jul. 2004, p. 178–181.
7. Suchanek F.M., Ifrim G. and Weikum G. Combining linguistic and statistical analysis to extract relations from web documents // KDD ’06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, p. 712–717.
8. Tsochantaridis I., Joachims T., Hofmann T., and Altun Y. Large margin methods for structured and interdependent output variables.

Рецензенты:

Сербулов Ю.С., д.т.н., профессор, проректор по научной работе Воронежского института высоких технологий, г. Воронеж.

Янсков А.И., к.т.н., начальник лаборатории ФГУП «Научно-исследовательский институт электронной техники».

Гаджиахмедов Н.Э., д.филос.н., профессор, зав. кафедрой теоретической и прикладной лингвистики Дагестанского государственного университета, г. Махачкала.
Работа получена 25.11.2011