УДК 004.8 + 004.032.26

ИЗВЛЕЧЕНИЕ ПРОСТЫХ ФАКТОВ ИЗ ТЕКСТА С ИСПОЛЬЗОВАНИЕМ РАСПРЕДЕЛЕННЫХ ПРЕДСТАВЛЕНИЙ

Слипченко С.В.
Международный научно-учебный центр информационных технологий и систем НАН и МОН Украины

Введение

Извлечение знаний из неструктурированных текстов остается одним из важнейших направлений исследований в области компьютерной лингвистики и искусственного интеллекта в целом. С развитием глобальной сети Internet и ростом объемов документов ручной поиск и обработка информации уже на текущий момент требуют значительных человеческих и финансовых ресурсов, покрывая при этом малую часть всей доступной информации, поэтому интерес к автоматическому извлечению знаний из текстов все увеличивается. Примерами знаний, которые представляют интерес, являются описания геополитических событий [1], результаты новейших исследований био-молекулярных структур [2], или отношения между объектами в текстах новостей [3]. При этом используются самые разнообразные модели и методы — от простых классификационных (например, сочетание CRF и правил Байеса [1], или SVM с использованием ядер на деревьях типизированных зависимостей [3]), до сложных вероятностных графических моделей (например, Markov Logic [2], которая сочетает в себе марковские сети и логику предикатов первого порядка).

Предлагаемый подход является развитием идей и методов рассуждений на примерах [4–7], отличительными особенностями которого является:

- использование базы примеров для построения выводов, что, в отличие от обучаемых моделей, позволяет легко интерпретировать процесс вывода;
- использование бинарных распределенных представлений, что, в отличие от логических и других не-распределенных моделей, позволяет быстро и эффективно находить примеры из базы, которые соответствуют входному образцу.

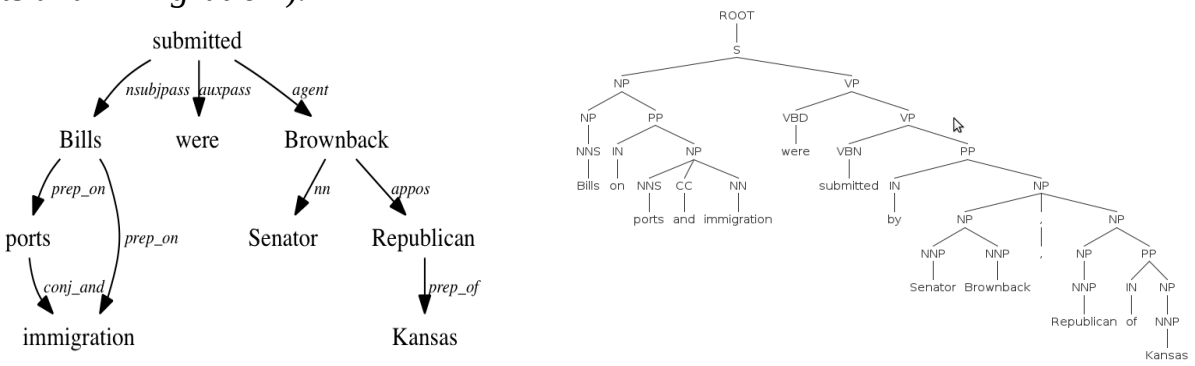
Методы на основе бинарных распределенных представлений успешно использовались для моделирования рассуждений по аналогии [6,7], однако при этом использовались предварительно подготовленные примеры высказываний [8,9]. Несмотря на достигнутые результаты по поиску, отображению и выводу по аналогии, именно вопрос формирования высказываний по текстовым описаниям примеров всегда вызывал горячий интерес.

Цель

Для развития методов обработки знаний на основе бинарных распределенных представлений необходимо разработать методы извлечения высказываний из текстовых описаний. При этом, как и в других работах [1–3], не ставиться задача синтаксического анализа текста, а в качестве основного инструмента используются методы поиска [6] и отображения [7] аналогов.

Результаты

В данной работе для синтаксического анализа текстов используется Стэнфордский вариант деревьев типизированных зависимостей [10], которые, в отличие от стандартных синтаксических деревьев, представляют грамматические отношения между словами, а не грамматическую структуру предложения (см. Рис. 1). Типизированные зависимости представляются тройками - *имя*, *управляющее* и *зависимое* слово. Для предложения **«Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas»** из примера на Рис. 1 мы получим зависимости: *nsubjpass*(*submitted*,*Bills*), *auxpass*(*submitted*,*were*), *agent*(*submitted*,*Brownback*), *nn*(*Brownback*,*Senator*), *appos*(*Brownback*,*Republican*), *prep_of*(*Republican*,*Kansas*), *prep_on*(*Bills*,*ports*), *conj_and*(*ports*,*immigration*) и *prep_on*(*Bills*,*immigration*). Преимущество деревьев типизированных зависимостей состоит в том, что используя полученные грамматические отношения эксперт может легко извлечь интересующее его отношение *submit*('*Brownback*', '*Bills on ports and immigration*').



а *б*

Рис. 1: Сравнение дерева типизированных зависимостей (а) и синтаксического дерева (б)

Для дерева типизированных зависимостей, которое строго говоря является направленным графом без циклов, с помощью процедуры контекстно-зависимого прореживания [11] можно сформировать распределенное представление предложения:

SENTENCE = <<*nsubjpass* $_{gov}$ ∨ submitted> ∨ <*nsubjpass* $_{dep}$ ∨ Bills>> ∨ … ∨

<<*agent* $_{gov}$ ∨ submitted> ∨ <agent $_{dep}$ ∨ Brownback>>

где *nsubjpass* $_{gov}$ и *nsubjpass* $_{dep}$ - случайные вектора ролей, ∨ - побитовая дизъюнкция векторов, □□ - операция контекстно-зависимого прореживания, а вектора Bills и Brownback сформированы аналогичным способом:

Bills = <<*prep_on* $_{gov}$ ∨ Bills> ∨ <*prep_on* $_{dep}$ ∨ ports>> ∨

<<*prep_on* $_{gov}$ ∨ Bills> ∨ <*prep_on* $_{dep}$ ∨ immigration>>

Brownback = <<*nn* $_{gov}$ ∨ Brownback> ∨ <*nn* $_{dep}$ ∨ Senator>> ∨

<<*appos* $_{gov}$ ∨ Brownback> ∨ < *appos* $_{dep}$ ∨ Kansas>>

Благодаря тому, что процедура контекстно-зависимого прореживания

$$<Z> = \vee_{k=1,K} (Z \wedge Z^{*}(k)),$$

где $Z^{*}(k)$ — случайная перестановка вектора Z, уменьшает число единиц в векторе и сохраняет информацию о порядке применения

$$<<A \vee B> \vee C> \neq <A \vee <B \vee C>>,$$

вектора схожих по структуре предложений - схожи, а различных - различны. Используя векторные представления входных предложений и обратное индексирование для предложений из базы по скалярному произведению векторов легко найти близкие по структуре предложения базы для последующего отображения и вывода.

Для отображения элементов входного предложения на элементы предложения из базы используется объединение представлений элементов с их ролями [12]:

Bills* = Bills ∨ *nsubjpass* $_{dep}$

Brownback* = Browback ∨ agent $_{dep}$

которое является некоторой обобщенной характеристикой всех путей от данного элемента к вершине дерева типизированных зависимостей. После определения по скалярному произведению векторов наилучших отображений между элементами выводы формируются простой заменой элементов из базы на соответствующие им входные.

Выводы

Предложенный метод извлечения знаний из текстов демонстрирует принципиальную возможность использования рассуждений по аналогии на основе распределенных представлений для решения этой задачи. Однако, несмотря на позитивные результаты первых экспериментов, требуется провести значительную работу по подбору параметров генерации и прореживания векторов с учетом лингвистических характеристик различных типов зависимостей и конкретных слов. Кроме того, необходима дополнительная адаптация метода для учета анафоры и катафоры.

Литература

1. Radinsky K. Mining the Web to Predict Future Events / K. Radinsky, E. Horvitz // Proc. of International Conference on Web Search and Data Mining. – Rome, Italy, 2013. – P. 255–264.
2. Poon H. Joint Inference for Knowledge Extraction from Biomedical Literature / H. Poon, L. Vanderwende // Proc. of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. – 2009. – P. 813–821.
3. Culotta A. Dependency Tree Kernels for Relation Extraction / A. Culotta, J. Sorensen // Proc. of the 42nd Annual Meeting on Association for Computational Linguistics. – Barcelona, Spain, 2004.
4. Амосов, Н М. Моделирование мышления и психики. – Киев: Наукова думка, 1965. – 304 с.
5. Амосов, Н М. Алгоритмы разума. Киев: – Наукова думка, 1979. – 223 с.
6. Rachkovskij D.A. Similarity-Based Retrieval with Structure Sensitive Sparse Binary Distributed Representations / D.A. Rachkovskij, S. V. Slipchenko // Computational Intelligence. – 2012. – Vol. 28, № 1. – P. 106 – 129.
7. Рачковский Д.А. Подходы к отображению аналогов с помощью распределенных представлений / Д.А. Рачковский, С.В. Слипченко // Компьютерная математика. – 2005. – № 1. – С. 55–69.
8. Forbus K.D. MAC/FAC: A model of similarity-based retrieval / K.D. Forbus, D. Gentner, K. Law // Cognitive Science. – 1995. – Vol. 19, № 2. – P. 141–205.
9. Falkenhainer B. The Structure-Mapping Engine: Algorithm and Examples / B. Falkenhainer , K.D. Forbus, D, Gentner // Artificial Intelligence. – 1989. – Vol. 41. – P. 1–63.
10. De Marneffe M.-C. The Stanford typed dependencies representation. / M.-C. De Marneffe, C.D. Manning // Proc. of COLING Workshop on Cross-framework and Cross-domain Parser Evaluation. – 2008. – P. 1–8.
11. Рачковский Д.А. Процедура связывания для бинарного распределенного представления данных / Д.А. Рачковский, С.В. Слипченко, Э.М. Куссуль,

Т.Н. Байдык // Кибернетика и системный анализ. – 2005. – № 3. – С. 3–18.
12. Слипченко С.В. Отображение и вывод по аналогии на основе нейросетевых распределенных представлений / С.В. Слипченко, Д.А. Рачковский // Proc. of Conference Knowledge - Dialogue - Solution. – Varna, Bulgaria, 2009. – Т. 9. – С. 95–101.