ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №1(24), 2015, с. 73–82

УДК 004.89:004.912

Е. А. Сулейманова, И. В. Трофимов

Референциальный профиль как инструмент для исследования референции в связном тексте

Аннотация. В статье предложена идея референциального профиля — инструмента для визуализации референции к некоторой сущности как процесса, разворачивающегося в связном тексте. Описаны формат и принципы разметки текста для построения референциального профиля.

Ключевые слова и фразы: референциальная разметка текста, визуализация разметки.

Введение

Референция в лингвистике — это отнесение текстового выражения к внеязыковому объекту. Сам внеязыковой объект, с которым соотносится текстовое выражение, называется его референтом. В контексте автоматического анализа естественного языка референциальные аспекты текста затрагиваются обычно в связи с задачей разрешения кореферентности — т.е. установления референциального тождества между текстовыми выражениями. Успешное разрешение кореферентности есть необходимое условие «понимания» текста на том уровне, который требуется от систем извлечения информации, поиска фактов в текстах, автоматического реферирования.

1. О разметке кореферентности

Разметка отношений кореферентности в текстах — активно развивающаяся область корпусной лингвистики. Известны несколько десятков проектов по созданию аннотированных корпусов с разметкой разных типов кореферентности для разных языков и разных жанров [1]. К самым масштабным инициативам, в рамках которых

Работа выполнена в рамках проекта РФФИ 14-07-00368 А «Исследование проблем и методов моделирования и использования общих знаний для разрешения кореферентности».

в разное время разрабатывались «золотые стандарты» разметки кореферентности, можно отнести MUC-7 [2], ACE [3] и многоязычный проект OntoNotes (в корпус включены тексты разных жанров на английском, китайском и арабском языках) [4]. Инициатива по созданию русскоязычного корпуса с разметкой кореферентности возникла в 2013 году в рамках национального форума по оценке систем автоматической обработки текста. Аннотированный корпус использовался для обучения и сравнительной оценки автоматических систем разрешения кореферентности на соревнованиях, проходивших во время Международной конференции по компьютерной лингвистике Диалог-2014 [5].

Создание аннотированного корпуса предполагает предварительную разработку принципов и правил разметки, которые будут положены в основу т. наз. аннотационной схемы (annotation scheme). Аннотационная схема регламентирует, что и каким образом будет размечаться в тексте. Обобщенная аннотационная схема для референциальной разметки включает три последовательных компонента:

- выбор аннотируемых элементов, или маркабул (от англ. markable) [6];
- разметка связей (природа размечаемых связей также определяется схемой разметки);
- разметка дополнительных признаков аннотируемых элементов.

Ввиду сложности и многообразия проявлений феномена кореферентности, референциальная разметка текста представляет собой очень непростую задачу. Даже если существенно ограничить для целей разметки множество рассматриваемых явлений (например, размечать только случаи конкретной определенной референции и только для узкого круга семантических категорий референтов), наряду с относительно благополучным «ядром» неизбежны периферийные ситуации, в которых далеко не очевидно, какое решение принять. Несмотря на попытки сформулировать некоторые универсальные межъязыковые принципы разметки кореферентности [7, 8] и разработать единую аннотационную схему [9], не существует общепринятого исчерпывающего стандарта референциальной разметки, который давал бы четкую инструкцию в любой ситуации и к тому же отвечал разным задачам.

2. Референциальный профиль

Референциальная разметка в нашем случае предназначена в первую очередь для удовлетворения исследовательских потребностей при решении задачи, которую в самом общем виде можно сформулировать как референциальный анализ текстов новостного жанра в системе извлечения информации. В частности, разметка способствует выявлению закономерностей выбора номинации целевых сущностей (на данном этапе исследования речь идет о лицах) для последующего моделирования процесса разрешения референции. Поскольку наш подход к разрешению референции строится вокруг понятия «референт» [10], а сам референт в явном виде в тексте не присутствует (наблюдению доступны лишь разбросанные по тексту упоминания о нем), у нас возникла идея визуализации референции к тому или иному референту как процесса, разворачивающегося в дискурсе. Сделать это можно, например, посредством построения по размеченному тексту графического объекта, наглядно представляющего содержащуюся в разметке информацию об упоминаниях некоторой внетекстовой сущности. Назовем такой объект референциальным профилем сущности. Референциальный профиль — это отображение последовательности текстовых упоминаний сущности на множество формализованных свойств упоминаний. Графически референциальный профиль сущности представляет собой точки-упоминания, часть из которых (а именно точки, соответствующие конкретно-референтным упоминаниям) соединены линией. Цвет точек и линий уникален для референта в рамках текста. Референциальный профиль строится автоматически на основании выполняемой вручную разметки текста.

3. Принципы разметки

Референциальная разметка основана на следующих основных принципах.

1. Разметка выходит за рамки собственно текста, поскольку соотносит текстовые упоминания с внетекстовыми объектами — моделями референтов (далее для краткости будем называть эти модели референтами). Референты явно строятся аннотатором в процессе разметки.

2. Разметке подлежат только упоминания сущностей заданных категорий.
3. Разметке подлежат не только конкретно-референтные упоминания целевых сущностей, но и прочие упоминания (предикатные, автонимные), которые потенциально могут впоследствии использоваться при повторной референции к ней (т.е. потенциальные антецеденты повторных конкретно-референтных упоминаний; ср. первичные и вторичные маркабулы в разметке RU-EVAL [5]).
4. Размечаемым элементом упоминания целевой сущности (маркабулой) является:
   - имя собственное; вершина именной группы — дескрипции;
   - местоимение (личные местоимения 3-го лица; притяжательные местоимения; относительные местоимения);
   - анафорический нуль (анафорический эллипсис именной группы).
5. В тех случаях, когда упоминание включает в себя вложенное упоминание (например, *его брат, адвокат задержанного*), между соответствующими маркабулами аннотируется синтаксическая связь. Отказ от Принципа максимального объема аннотируемых элементов [8] позволяет более наглядно представлять случаи именования одного референта через отношение к другому.
6. Местоимения 1-го и 2-го лица не размечаются, в том числе и в составе прямой речи.
7. Прямая речь размечается на общих основаниях. При этом фрагменты, соответствующие прямой речи, при разметке выделяются (предусмотрен специальный инструмент для маркировки границ прямой речи). На графике референциальных профилей области, соответствующие прямой речи, будут выделены фоном.
8. Границы абзацев подлежат разметке (это делается автоматически на основании формата исходного текста). На графике границам абзацев соответствуют вертикальные пунктирные линии.
9. При разметке упоминаниям приписываются значения определенных признаков. Эти значения отображаются на референциальном профиле.

3.1. Признаки упоминаний

Признак «Тип упоминания». Имеет значения:

- «дескрипция + имя собственное»;
- «имя собственное»;
- «дескрипция»;
- «местоимение»;
- «анафорический нуль».

Значение этого признака определяет координату точки-упоминания по оси ординат. Замечание: упоминания типа «дескрипция + имя собственное» размечаются особым образом, что позволяет отразить структуру упоминания и выделить входящие в его состав элементы (см. приведенный далее пример).

Признак «Категория упоминания». Имеет значения:

- конкретно-референтное;
- свойство/аспект референта (это значение приписывается всем вторичным маркабулам).

Упоминаниям первой категории соответствуют заштрихованные точки, упоминаниям второй категории — незаштрихованные.

Следующие бинарные признаки могут характеризовать только упоминания типа «дескрипция» (в том числе в составе упоминания типа «дескрипция + имя собственное»).

Признаком «реляционная дескрипция» помечается вершина упоминания лица, подчиняющая по реляции другое упоминание лица [11]. На профиле признак отображается в виде буквы R, вписанной в точку-упоминание, а само отношение-реляция — в виде стрелки между упоминаниями. Признак «ситуативно обусловленная дескрипция» введен для наглядного выделения случаев выбора контекстно обусловленного референциального выражения (например, *задержанного* в приведенном далее примере). На профиле такие упоминания будут помечены буквой S.

Признак «титул или аналог» (отображаемый в виде буквы T) полезен для выделения случаев т. наз. этикетного употребления дескрипций (*тетя Маша, президент Обама*).

4. Примеры

На рис. 1 приведен пример референциальной разметки текста новости. Разметка выполнена в инструментальной среде ИСИДА-Т [12].

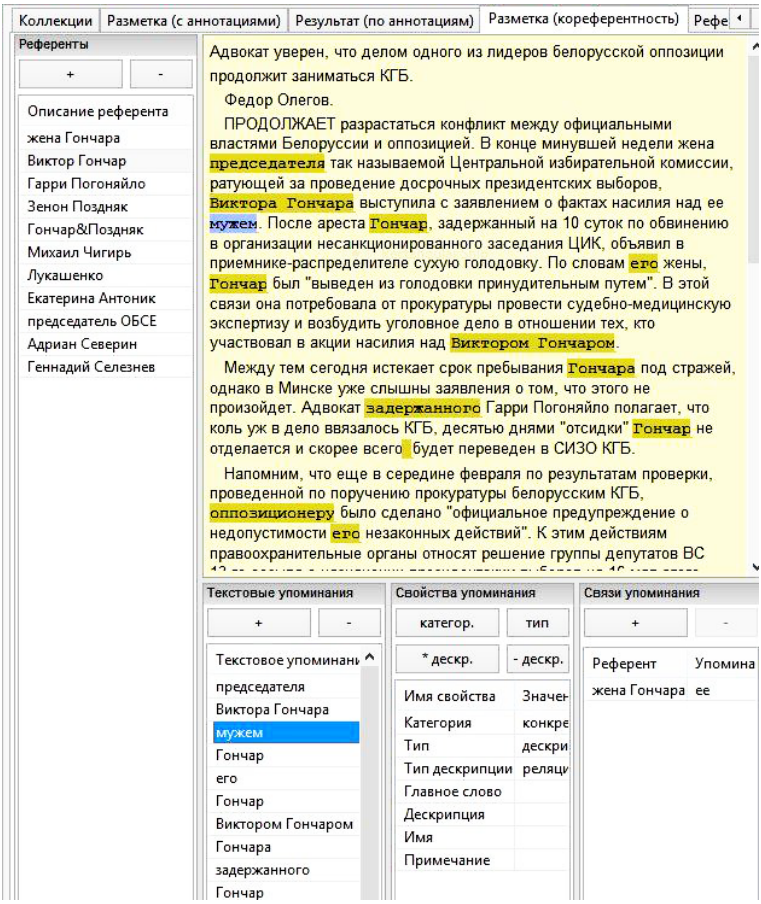

Рис. 1. Фрагмент разметки текста для построения референциальных профилей

Референциальные профили, построенные по размеченному фрагменту текста, изображены на рис. 2 (изображение повернуто на 90 градусов).

Заключение

Референциальный профиль, задуманный как некоторый способ визуализации референции, может быть полезен не только в качестве

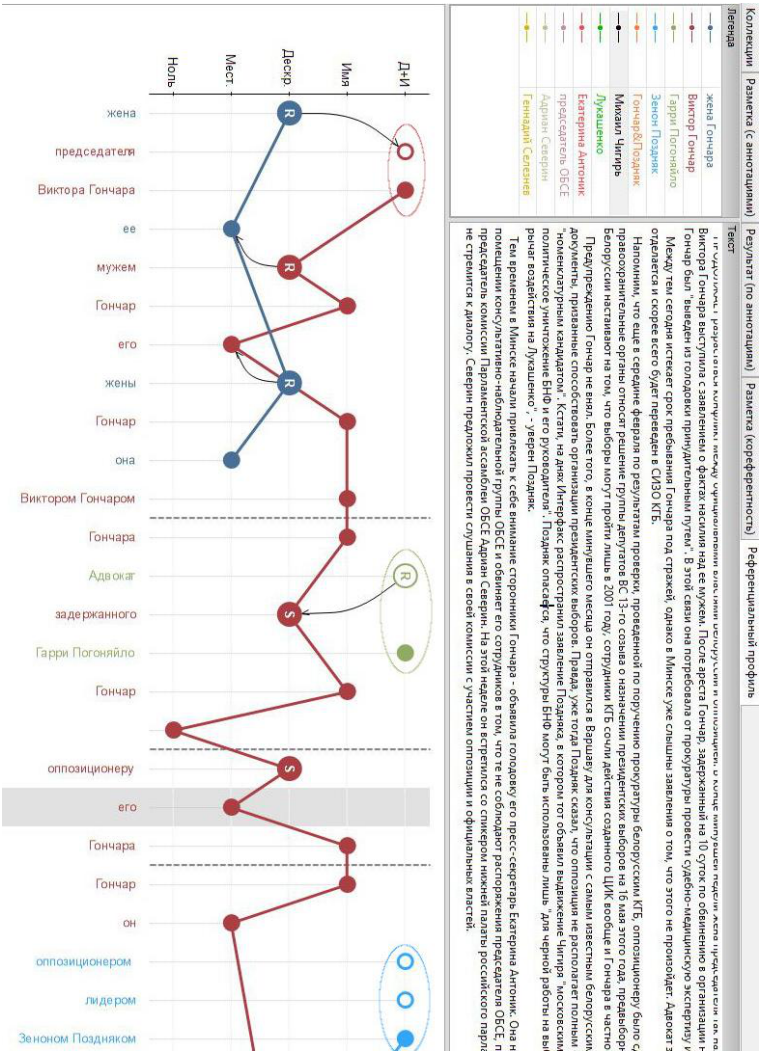

Рис. 2. Референциальные профили, построенные по разметке

инструмента аналитической деятельности исследователя. Разметка коллекции текстов — трудоемкий процесс, требующий постоянной концентрации внимания. Референциальный профиль, который строится по аннотируемому тексту непосредственно в процессе разметки,

служит для аннотатора удобным наглядным средством контроля за ее качеством и снижает напряженность работы.

Список литературы

[1] А. Ю. Недолужко, «Кореферентные отношения в тексте — сравнительный анализ размеченных данных», По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.), Компьютерная лингвистика и интеллектуальные технологии, т. **9(16)**, Изд-во РГГУ, М., 2010 ↑ 73.

[2] L. Hirschman. “MUC-7 Coreference Task Definition. Version 3.0”, Proceedings of the Seventh Message Understanding Conference (Fairfax, 1998) ↑ 74.

[3] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, R. M. Weischedel, “The Automatic Content Extraction (ACE) Program — Tasks, Data, and Evaluation”, *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, LREC’2004 (European Language Resources Association (ELRA), Lisbon, Portugal, May 2004) ↑ 74.

[4] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, Yu. Zhang. “CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes”, Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task (Jeju Island, Korea, 13 July 2012), pp. 1–40 ↑ 74.

[5] S. Toldova, A. Roytberg, A. Ladygina, M. Vasilyeva, I. Azerkovich, M. Kurzukov, G. Sim, D. Gorshkov, A. Ivanova, A. Nedoluzhko, Y. Grishina, «RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian», По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.), Компьютерная лингвистика и интеллектуальные технологии, т. **13(20)**, Изд-во РГГУ, М., 2014 ↑ 74, 76.

[6] А. А. Кибрик, Г. Б. Добров, Д. А. Залманов, А. С. Линник, Н. В. Лукашевич, «Референциальный выбор как многофакторный вероятностный процесс», По материалам международной конференции «Диалог 2010», Компьютерная лингвистика и интеллектуальные технологии, т. **9(16)**, Изд-во РГГУ, М., 2010, с. 173–180 ↑ 74.

[7] O. Krasavina, C. Chiarcos. “PoCoS — Potsdam Coreference Scheme”, Proceedings of the Linguistic Annotation Workshop (Prague, June 2007), pp. 156–163 ↑ 74.

[8] О. Н. Красавина. *Корпусно-ориентированное исследование референции (принципы аннотации и анализ данных)*, Дис. ... канд. филол. наук, М., 2006 ↑ 74, 76.

[9] L. Hasler, C. Orasan, K. Naumann. “NPs for Events: Experiments in Coreference Annotation”, Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06) (Genoa, Italy, 2006), pp. 1167–1172 ↑ 74.

[10] Е. А. Сулейманова, И. В. Трофимов. «О подходе к отождествлению сущностей в рамках задачи извлечения информации из текстов», *Программные системы: теория и приложения*, **4**:1(15) (2013), с. 15–30, URL http://psta.psiras.ru/read/psta2013_1_15-30.pdf ↑ 75.

[11] Е. А. Сулейманова. «О комплексном подходе к разрешению реляционно-аппозитивных неоднозначностей», *Программные системы: теория и приложения*, **5**:4(22) (2014), с. 41–46, URL http://psta.psiras.ru/read/psta2014_4_41-66.pdf ↑ 77.

[12] Д. А. Кормалев, Е. П. Куршев, Е. А. Сулейманова, И. В. Трофимов. «Технология извлечения информации из текстов, основанная на знаниях», *Программные продукты и системы*, 2009, №2(86), с. 62–66 ↑ 77.

Рекомендовал к публикации *к.т.н. Е. П. Куршев*

Об авторах:

Елена Анатольевна Сулейманова

Научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, одна из разработчиков технологии построения систем извлечения информации.

e-mail: yes@helen.botik.ru

Игорь Владимирович Трофимов

Старший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна РАН, специалист по технологиям автоматической обработки естественного языка, извлечения информации, автоматического планирования.

e-mail: igor@warlock-98.botik.ru

Пример ссылки на эту публикацию:

Е. А. Сулейманова, И. В. Трофимов. «Референциальный профиль как инструмент для исследования референции в связном тексте», *Программные системы: теория и приложения*, 2015, **6**:1(24), с. 73–82.

URL http://psta.psiras.ru/read/psta2015_1_73-82.pdf

Elena Suleymanova, Igor Trofimov. *Referential profile — a tool for studying reference in discourse.*

Abstract. The paper suggests the idea of referential profile as a tool for visualizing the process of referencing an entity throughout the discourse. The referential profile is automatically constructed from manually annotated text. Annotation principles and guidelines are outlined. (*In Russian*).

Key Words and Phrases: coreference markup, markup visualization.

Sample citation of this publication

Elena Suleymanova, Igor Trofimov "Referential profile — a tool for studying reference in discourse", *Program systems: theory and applications*, 2015, **6**:1(24), pp. 73–82. (*In Russian.*)

URL http://psta.psiras.ru/read/psta2015_1_73-82.pdf