Открытый корпус для разрешения кореферентности для русского языка

Егор Будников, Дарья Зверева
ABBYY, МФТИ
Дарья Максимова
НИУ ВШЭ

ABBYY

План доклада

- Постановка задачи
- Характеристики корпуса
- Модель разметки
- Текущие результаты и дальнейшие действия

Постановка задачи

Задача разрешения кореферентности в тексте заключается в определении того, относятся ли два *упоминания* к одной *сущности*.

Упоминание – фраза, ссылающаяся на объект или явление.

Сущность – непосредственно сам объект или явление.

Пример:

Существующие корпуса

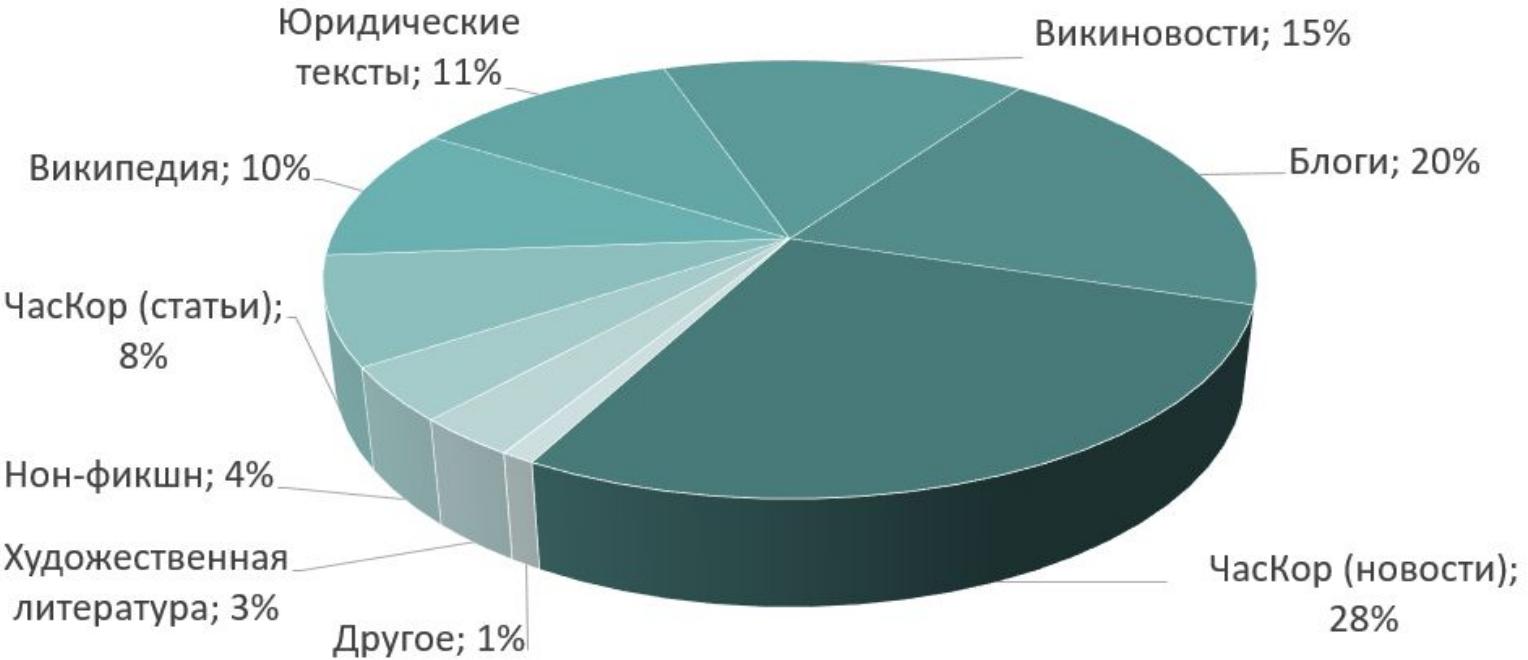
- Message Understanding Conference-6 [Grishman, Sundheim 1996]
  - Английский (318 документов)
- CoNLL-2012 Shared Task [Pradhan et al. 2012]
  - Английский (2384 документа), Арабский (447), Китайский (1729)
- Prague Dependency Treebank [Nedoluzhko 2016]
  - Английский, Чешский (50к предложений, >60к связей)
- RuCor [Toldova et al. 2014]
  - Русский (181 документ)

ABBYY

Новый корпус

- Источник текстов: Открытый корпус русского языка (OpenCorpora.org)
- Размер: 3729 документов

Новый корпус: распределение документов



ABBYY

Разметка нового корпуса

- Слой упоминаний
- Слой кореференциальных цепочек
- Морфологический слой
- Семантико-синтаксический слой
- Эмбеддинги семантических классов

Разметка слоя упоминаний

Источники:

- ABBYY Compreno (автоматическая)
- Разметчики (эталонная)

Формат:

- Идентификатор упоминания
- Смещение упоминания
- Длина упоминания

ABBYY

Разметка слоя упоминаний

Размечаются:

- Персоны, Локации, Организации, иные именные сущности

  Ключевое слово + Идентификатор. Всегда имеют референта

- Именные группы

  Реальные объекты или абстрактные понятия, на которые ссылаются позднее

- Местоимения и местоименные группы

  Те, которые могут иметь референта (кроме негативных, взаимных и возвратных)

Разметка слоя упоминаний: интересные случаи

ABBYY

- Возвратные и взаимные местоимения

  Вася сходил в магазин. Мальчик купил *<u>*себе*</u> утюг.

  Коля и Катя во всем *<u>*друг друга*</u> поддерживают.

- Синонимичные названия (псевдонимы)

  107-мм пушка образца 1940 года (<u>*М-60*</u>)

- “Прилагательные”, имеющие референтов

  <u>*Британская*</u> разведка; <u>*Североамериканская*</u> литература; <u>*мой*</u> дом; <u>*который*</u> смог

Разметка слоя упоминаний: интересные случаи [2]

ABBYY

- Именованные упоминания *vs.* Неименованные упоминания

  Маленький *мальчик Вася*, *который* умеет плавать (2 упоминания)

  Маленький мальчик, который умеет плавать (1 упоминание)

- Описательные именные группы

  Вася признан *самым смелым мальчиком в классе*. *Самый смелый мальчик* на прошлой неделе снял котенка с дерева.

- Описательные именные группы [2]

  Лидером мнений на этой неделе оказался *Петя*. На прошлой неделе *им* был *Вася*.

ABBYY

Разметка слоя цепочек

Источник:

- Разметчики (эталонная)

Формат:

- Идентификатор упоминания
- Смещение упоминания
- Длина упоминания
- Идентификатор цепочки

Разметка слоя цепочек: интересные случаи

- Часть и целое

  Петя$_1$ и *Вася*$_2$ одноклассники. *Они*$_3$ каждый день ходят в школу вместе. *Мальчики*$_3$ живут в соседних подъездах. *Петя*$_1$ живет на третьем этаже, а *Вася*$_2$ на пятом.

- Описательные именные группы

  Лидером мнений$_1$ на этой неделе оказался Петя. На прошлой неделе *им*$_1$ был Вася.

ABBYY

Морфологический слой

Источник:

- OpenCorpora (эталонная)

Формат:

- Идентификатор токена
- Смещение токена
- Длина токена
- Лемма
- Морфологические теги

Семантико-синтаксический слой

Источник:

- ABBYY Compreno (автоматическая)

Формат:

- Смещение токена
- Токен
- Смещение родительского токена
- Идентификатор семантического класса
- Идентификатор синтаксической парадигмы

Эмбеддинги семантических классов

Источник:

- ABBYY Compreno (обучено на 800M слов)

Формат:

- Идентификатор семантического класса
- Вектор размерности 200

ABBYY

Статус работ

- Разметок упоминаниями хотя бы одним разметчиком — 432 документа
- Разметок упоминаниями двумя разметчиками — 183 текстов
- Сведено разметок упоминаниями — 120 текстов
- Разметок кореференциальными цепочками — 120 текстов

Текущий размеченный корпус лежит тут: https://goo.gl/U2vTvS

ABBYY

Статус работ

	Сейчас	*Август 2018*	*Декабрь 2018*	*Однажды*
Текстов	120	400	800	3700
Упоминаний	8700	29K	58K	250K
Упоминаний в кореф. цепочках	6500	22K	44K	200K
Кореф. цепочек	1300	4K	8K	35K

ABBYY

Мера согласия разметок упоминаниями

	Полнота	*Точность*	*F-мера*
Верификаторы 1 *vs.* 2	76.40%	68.50%	72.20%
Верификаторы 1,2 *vs.* финальная разметка упоминаниями	77.40%	93.70%	84.80%
Авторазметка *vs.* финальная разметка упоминаниями	94.30%	30.40%	45.90%

Список литературы

1. Anisimovich et al. 2012 — Anisimovich K., Druzhkin K., Minlos F., Petrova M., Selegey V., and Zuev K. (2012). Syntactic and semantic parser based on ABBYY Compreno linguistic technologies. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference ”Dialogue”, vol. 11, pp. 91– 103.

2. Grishman, Sundheim 1996 — Grishman, R., & Sundheim, B. (1996). Message understanding conference-6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics (Vol. 1).

3. Korobov 2015 — Korobov M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: Khachay M., Konstantinova N., Panchenko A., Ignatov D., Labunets V. (eds) Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science, vol 542.

4. Luo 2005 — Luo, X. (2005, October). On coreference resolution performance metrics. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 25-32). Association for Computational Linguistics.

5. Moosavi, Strube 2016 — Moosavi, N. S., & Strube, M. (2016). Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. In ACL (1)

6. Nedoluzhko 2016 — Ngụy Giang Linh, Michal Novak, Anna Nedoluzhko (2016). Coreference Resolution in the Prague Dependency Treebank. (ÚFAL/CKL Technical Report #TR-2011-43). Prague: Universitas Carolina Pragensis.

7. Pradhan et al. 2012 — Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012, July). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In Joint Conference on EMNLP and CoNLL-Shared Task (pp. 1-40). Association for Computational Linguistics.

8. Stepanova et al. 2016 — Stepanova M. E., Budnikov E. A., Chelombeeva A. N., Matavina P. V., Skorinkin D. A. (2016). Information Extraction Based on Deep Syntactic-Semantic Analysis. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference ”Dialogue”, pp. 721-732.

9. Toldova et al. 2014 — Toldova, S., Roytberg, A., Ladygina, A., Vasilyeva, M., Azerkovich, I., Kurzukov, M., ... & Grishina, Y. (2014). RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian. Computational Linguistics and Intellectual Technologies, 13(20), 681-694.

10. Vilain et al, 1995 — Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995, November). A model-theoretic coreference scoring scheme. In Proceedings of the 6th conference on Message understanding (pp. 45-52). Association for Computational Linguistics.


THANK YOU