Кореферентные отношения в тексте — сравнительный анализ размеченных данных

Coreferential relations in the text — a comparative analysis of annotated data

Недолужко А. Ю. (nedoluzko@ufal.mff.cuni.cz)

Карлов университет, Прага, Чехия

Данная работа опирается на находящийся в разработке проект разметки именной кореференции и ассоциативной анафоры на материале синтаксически аннотированного корпуса чешских текстов PDT 2.0. В ходе работы над разметкой кореференции выяснилось, что относительно низкое соответствие между разметчиками объясняется не столько их ошибками и невнимательностью, сколько объективной неоднозначностью интерпретации текста. В докладе приводится классификация типов и возможных причин несоответветствий и рассматриваются некоторые типовые примеры множественности интерпретаций кореферентных отношений в тексте.

1. Введение

Одним из наиболее актуальных направлений в области обработки естественного языка является в настоящее время извлечение информации из текста. Выявление кореферентных и анафорических отношений из связного текста — необходимый механизм для решения этой задачи.

Проект разметки именной кореференции и ассоциативной анафоры на относительно обширном корпусном материале чешских текстов PDT 2.0 (Prague Dependency Treebank) — один из нескольких десятков мировых корпусных исследований анафорических отношений (напр. Hirschman 1998, Poesio 2004a для английского языка, Recasens — Martí 2010 для испанского и английского, Poesio 2004б для итальянского, Krasavina — Chiarcos 2007 для английского и немецкого и др.). Разметка именной кореференции и ассоциативной анафоры на PDT 2.0 производится вручную и частично автоматически на глубинно-синтаксическом уровне синтаксически аннотированного корпуса чешских публицистических текстов (подробнее с проектом синтаксической разметки можно ознакомиться в Hajičová 2006, Недолужко 2008). В настоящее время размечено около половины всего корпуса PDT 2.0. Регулярно проводятся измерения соответствия между разметками разных аннотаторов (разметчиков) и составляется типология несоответствий, что позволяет делать первые выводы о возможных причинах этих несоответствий. Классификации несоответствий между разметчиками и попытке объяснения некоторых причин их возникновения посвящена данная работа.

2. Типология отношений кореференции и ассоциативной анафоры, размечаемых на PDT 2.0

В PDT 2.0 представлена разметка трех типов кореферентных и анафорических отношений[2]:

1. грамматическая кореференция, где антецедент высчитывается на основе грамматических правил языка. К грамматической кореференции относятся, например, кореференция возвратных местоимений[3], относительных местоимений (*человек, который пьет*), актантов в реципрокальных конструкциях и т. д. Грамматическая кореференция практически никогда не переходит границ предложения

[1] Эта работа была поддержана грантами GACR 405/09/0729 и GAUK 4383/2009.

[2] К более подробному описанию типов см напр. (Nedoluzhko 2007, Nedoluzhko-Mirovsky-Pajas 2009)

[3] В чешском языке возвратное местоимение «ся» всегда является отдельной лексемой (клитикой).

Таблица 1. PDT 2.0 — статистические данные

количество предложений в PDT 2.0	115 844
количество документов в PDT 2.0	7110
размечено грамматической кореференции	100,00 %
размечено прономинальной текстовой кореференции	100,00 %
размечено именной текстовой кореференции	50,00 %
размечено ассоциативной анафоры	50,00 %
количество узлов, связанных кореференцией и ассоциативной анафорой	67 071
процент узлов, связанных кореференцией и ассоциативной анафорой	20,45 %

2. т. наз. текстовая кореференция, где соотношение между кореферентными членами реализуется не только за счет грамматических средств языка, но и на основании знания контекста. Текстовая кореференция может легко переходить границы предложения. Различается прономинальная и именная текстовая кореференция[4].

- прономинальная текстовая кореференция размечается в случае, если в качестве второго члена кореферентного отношения выступают личные и притяжательные местоимения третьего лица, указательное местоимение *этот* в субстантивной функции или эллиптированное и восстановленное на глубинно-синтаксическом уровне местоимение 3-го лица[5]. На глубинно-синтаксическом уровне в PDT эти местоимения восстанавливаются, и им присваивается тектограмматическая лемма #PersPron.
- разметка именной текстовой кореференции является продолжением предшествующей ей разметки прономинальной кореференции. В качестве второго члена кореферентного отношения выступают в основном имена существительные и некоторые наречия *(там, тогда* и др.*)*. В некоторых случаях в отношении кореферентности могут участвовать прилагательные (притяжательные прилагательные, прилагательные, образованные от имен собственных и др.) и числительные (выступающие в субстантивной функции и релевантные для связности текста). При разметке именной текстовой кореференции используется два типа отношений — отношение между конкретнореферентными ИГ (дефолтный тип 0) и отношения между нереферентными и родовыми ИГ (тип NR), причем тип отношения определяется по второму члену анафорического отношения[6].

3. ассоциативная анафора (bridging anaphora), где анафорический член и антецедент уже не кореферентны, но между ними имеется семантическое отношение определенного типа. В настоящее время размечается шесть типов отношений: часть — целое (напр. *Бавария — Германия*), множество — подмножество/элемент множества (*студенты — три студента*), отношение дискурсивного контраста (*Люди не жуют, жуют только коровы*), отношение объекта и его функции/позиции (напр. *школа — учитель*), эсплицитное анафорическое отношение между некореферентными членами (*учителя — такие же учителя*) и отношение «остальное». Последний тип размечается у отношений типа место — житель (*Москва — москвич*), автор — творение, вещь — хозяин, у родственных отношений (*дед — внук*), а также у некоторых предикатно-аргументных отношений (*предпринимательство — предприниматель, спор — участник конфликта* и др.).

Грамматическая и прономинальная текстовая кореференция обработаны полностью на всем корпусе PDT 2.0[7]и в данном докладе учитываются только в статистических данных. Именная кореференция и ассоциативная анафора размечены на половине этого корпуса и являются предметом анализа в данной работе. Некоторые статистические данные о величине корпуса PDT 2.0 и количестве размеченных на нем кореферентных и анафорических отношений представлены в таблице 1.

---

[4] В случае прономинальной текстовой кореференции речь идет, как правило, и об анафорическом отношении. Для именной кореференции размечается именно соответствие референтов данного отношения без учета того, является ли это отношение также анафорическим.

[5] Являясь языком pro-drop, чешский язык имеет сильную тенденцию опускать личные местоимения в анафорических конструкциях (напр. чеш. 0 Nechtěl to říkat. vs. рус. Он не хотел этого говорить.)

[6] Различие ИГ на конкретнореферентные и родовые является свойством независимым от участия данной ИГ в кореферентном отношении. Однако, не имея технической возможности приписывать данный признак всем именным группам, мы ограничиваемся разметкой этого различия только у ИГ, вступающих в кореферентные отношения.

[7] См. Kučová L. и др. 2003

Таблица 2. Соотношение типов именной текстовой кореференции и ассоциативной анафоры

отношение		**количество узлов**	**%**
грамматическая кореференция		11 327	17,54
текстовая прономинальная кореференция		10 747	16,64
текстовая именная кореференция	тип 0	12 034	18,63
	тип NR	2740	4,24
	всего	14 774	22,87
ассоциативная анафора	множество — подмножество/элемент	3307	5,12
	часть — целое	1408	2,18
	дискурсивный контраст	655	1,01
	объект — функция/позиция	355	0,55
	некореферентная анафора	24*	0,04
	остальное	733	1,13
	всего	6482	10,04

* Малое количество отношение типа ANAF объясняется тем, что данный тип стал размечаться только на последнем этапе аннотирования. В ближайшем будущем планируется вернуться к той части корпуса, где этот тип размечен не был, и пройти его еще раз.

Соотношение типов именной текстовой кореференции и ассоциативной анафоры представлено в таблице 2.

3. Измерение соответствий между разметчиками

Аннотирование именной кореференции и ассоциативной анафоры проводится двумя разметчиками с лингвистическим образованием, причем на данном этапе разметка производится «в один слой», т. е. разметчики работают на разных текстах. Тем не менее мы регулярно проверяем соответствие между разметчиками на небольших порциях текстов. В таблице 3 рассмотрено шесть измерений соответствий разметок у разметчиков А и Б, проведенных с приблизительно двухмесячным интервалом. Для измерения соответствия при выборе антецедента как для текстовой кореференции, так и для ассоциативной анафоры, мы использовали F1-measure (Chinchor 1992). Соответствие типов отношений на совпавших парах посчитано в процентах.

Как видно из таблицы 3, успешность разметки не имеет тенденции постоянно возрастать, как бы мы того ожидали. Более того, два первые соответствия имеют по многим параметрам бо́льшую успешность, чем последующие измерения 3–5. Успешность последнего шестого измерения вновь несколько возрастает. Тем не менее F1-measure для текстовой кореференции ни в одном измерении не превышает 75,2 %, т. е. для применения наших данных при автоматическом обучении, тестировании и оценке автоматической разметки новых данных их качества еще недостаточно. Еще меньшим является соответствие между разметчиками при установлении отношений ассоциативной анафоры (F1-measure не более 55,5 %). Однако интересно заметить, что при совпадении узлов первого и второго членов отношения кореференции или ассоциативной анафоры (т. е. при наличии одинаковой

Таблица 3. Соответствие разметок у разметчиков А и Б

	1-е изм.	**2-е изм.**	**3-е изм.**	**4-е изм.**	**5-е изм.**	**6-е изм.**
кол-во текстов	3	1	1	2	3	8
кол-во предложений	41	40	101	106	100	211
TKR*, А=Б при выборе антецедента	77,2	63,3	65,6	68,6	62	75,2
TKR, А=Б при выборе антецедента и типа отношения	65,9	39,6	54,5	64,7	50	64,4
TKR, только типы	85,2	62,5	83	94,3	80,5	85,6
bridging**, А=Б при выборе антецедента	55,5	31	35,4	42,2	40,8	43,5
bridging, А=Б при выборе антецедента и типа отношения	55,5	31	33,9	39,1	30	40,9
bridging, только типы	100	100	94,1	92,8	71,4	96,1

* TKR = именная текстовая кореференция.
** Bridging = ассоциативная анафора.

стрелки) совпадение между разметчиками на типе отношений уже достаточно велико (в среднем более 90 %), что свидетельствует о том, что низкая степень совпадений **не** обуславливается слишком сложной типологией размечаемых отношений.

4. К вопросу надежности измерения соответствий между разметчиками

Данные таблицы 3 не представляют стандартной статистической ценности, так как соответствие между разметчиками измерялось на разном количестве предложений различного уровня сложности (см. строки «кол-во текстов» и «кол-во предложений» в таблице 3). Тем не менее они позволяют сделать несколько теоретических наблюдений.

Сравнение параллельных разметок показывает, что **степень соответствия между разметками в высшей степени зависит от величины и сложности текста**. Чем короче текст, тем яснее отношения между его членами, тем выше соответствие между разметками разных аннотаторов. Чем больше в тексте абстрактных понятий, именных групп с родовым денотативным статусом, тем больше вероятность различной интерпретации кореферентных отношений и тем соответствие между разметками ниже. Так для первого измерения использовались три текста длиной не более 15 предложений (см. данные в таблице 3), содержащих в основном только конкретнореферентные ИГ, в результате чего соответствие между разметчиками оказалось достаточно высоким (F1-measure=77,2 % при выборе антецедента и 65,9 % с учетом соответствия при выборе типа на совпавшем отношении). Для второго измерения использовался только один текст, который был однако существенно длиннее предыдущих (40 предложений) и который содержал большое количество родовых и абстрактных понятий. В результате соответствие при выборе антецедента для текстовой кореференции упало до F1-measure=63,3 %, а совпадение на ассоциативной анафоре оказалось совсем низким (F1-measure=31 %). Наиболее надежным является последнее шестое измерение, где для оценки степени соответствия между разметками разных аннотаторов использовалось 8 текстов (211 предложений) различной длины и степени сложности.

5. Типология несоответствий при разметке кореференции и ассоциативной анафоры

Рассмотрим более подробно соответствие между разметками аннотаторов А и Б при последнем (шестом) измерении, причем сосредоточимся только на определении элементов отношений кореференции и ассоциативной анафоры без учета их дальнейшей типологии. Получаем несоответствия трех следующих типов:

- Разметчик А отметил отношение кореференции/ассоциативной анафоры там, где разметчик Б его не увидел, см (1) — в 69 % случаев.

(67) чеш. *Na této stránce vám budeme představovat jednotlivé obory <u>národního</u> hospodářství.* […] *Bylo to v době, kdy se nebývale zvýšil zájem zahraničních turistů a podnikatelů o návštěvu <u>České republiky</u>.*
рус. *На этом сайте будут представлены отдельные отрасли <u>национальной</u> экономики*. […] *Это было в тот период, когда небывало возрос интерес иностранных туристов и предпринимателей к посещению <u>Чешской Республики</u>.*

текстовая кореференция между *národní (национальной)* и *České republiky (Чешской Республики)*):

- разметчик А: отметил отношение кореференции,
- разметчик Б: не отметил это отношение.
- Различный выбор первого или второго члена отношения, см (2) и рис. 1 — в 20 % случаев;

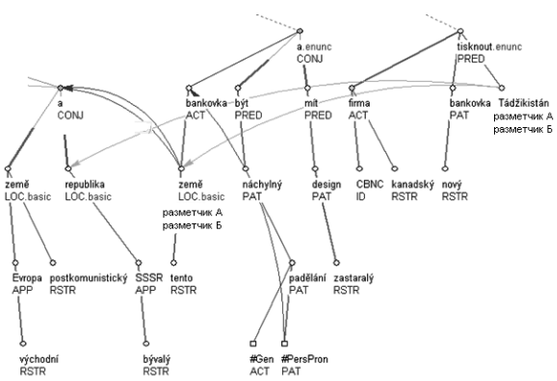

Рис. 1. Различный выбор первого члена отношения

(68) чеш. *Tiskárny bankovek mají i nové zákazníky, především <u>v postkomunistických zemích východní Evropy a republikách bývalého SSSR</u>. Bankovky <u>v těchto zemích</u> jsou náchylné na padělání a mají zastaralý design. Kanadská firma CBNC bude tisknout nové bankovky pro <u>Tádžikistán</u>*
рус. *У монетных дворов есть и другие клиенты, прежде всего <u>в посткоммунистических государствах Восточной Европы и в республиках бывшего СССР</u>. Банкноты <u>в этих странах</u> легко подделать, и у них устаревший дизайн. Канадская фирма CBNC будет печатать новые банкноты для <u>Таджикистана</u>.*

ассоциативная анафора типа «множество — подмножество»:

- разметчик А: отметил отношение «*Таджикистан*» на «*республиках бывшего СССР*»,
- разметчик Б: отметил отношение «*Таджикистан*» на «*в этих странах*», т. е. на всю сочинительную конструкцию «*в посткоммунистических государствах Восточной Европы и в республиках бывшего СССР*».

- Разметчик А отметил отношение кореференции там, где разметчик Б отметил отношение ассоциативной анафоры, см (3) — в 11 % случаев.

(69) чеш. *I přes klesající inflaci ve světě, a tedy nižší potřebu peněz v oběhu, je tisk bankovek a výroba bankovkového papíru jedním z nejlukrativnějších odvětví.* […] … *Rozšíření bankovních automatů vyžaduje neustálý přísun nepoškozených bankovek*.
рус. *Несмотря на снижение инфляции в мире, и соответственно меньшую потребность в оборотных денежных средствах, печать банкнот и производство специальной бумаги является одной из наиболее доходных отраслей.* […] … *В связи с расширением сети банкоматов требуется постоянное пополнение неповрежденных банкнот*.

отношение между *bankovek (банкнот) — nepoškozených bankovek (неповрежденных банкнот)*

- разметчик А: отметил отношение текстовой кореференции типа NR;
- разметчик Б: отметил отношение ассоциативной анафоры типа «множество — подмножество»

6. Причины расхождений разметок разных аннотаторов

Анализ причин расхождений разметчиков при аннотации кореференции и ассоциативной анафоры показал очень интересные результаты. Наиболее вероятной причиной половины расхождений является неоднозначность интерпретации текста. Так например в (3) нельзя точно сказать, кто из разметчиков более прав — тот, кто отметил отношение между *bankovek (банкнот) — nepoškozených bankovek (неповрежденных банкнот)* как ассоциативную анафору типа «множество-подмножество» (т. е. неповрежденные банкноты подмножество всех изготавливаемых банкнот) или тот, кто обозначил отношение между этими именными группами как кореферентное (ведь в сущности все изготавливаемые и выпускаемые в оборот банкноты являются неповрежденными). Подобная ситуация имеет место в примере (2).

Еще в 23 % разметчики расходятся в глубине интерпретации анафорических отношений в тексте, т. е. один из разметчиков отмечает отношение там, где второй считает его уже слишком расплывчатым, что по сути тоже можно рассматривать как неоднозначность интерпретации. Так в (4) отношение между *cestovní ruch* (*туризм*), *hotelových kapacit* (*мест в гостиницах*) и *zahraničních turistů a podnikatelů* (*иностранных туристов и предпринимателей*) может быть интерпретировано как ассоциативная анафора (разметчик А) или как отношение более расплывчатого свойства, которое находится уже за границами разметки (разметчик Б).

(70) чеш. *Méně výnosný cestovní ruch. Hotelových kapacit je mnohem víc než současná poptávka*. […] *Bylo to v době, kdy se nebývale zvýšil zájem zahraničních turistů a podnikatelů o návštěvu České republiky, především Prahy*.
рус. *Менее доходным является туризм. Количество мест в гостиницах существенно превышает современный спрос*. […] *Это было в тот период, когда небывало возрос интерес иностранных туристов и предпринимателей к посещению Чешской Республики*.

Примерно четверть несоответствий может быть интерпретировано как ошибка разметчика (см пример (1)).

Отдельные несоответствия (4 %) являются следствием неточно сформулированных правил разметки.

Таблица 4 обобщает рассмотренные данные.

Таблица 4. Причины расхождений разметок разных аннотаторов

неоднозначность интерпретации	69,00 %
глубина интерпретации	23,00 %
ошибка разметчика	23,00 %
неточность правил разметки	4,00 %

7. Выводы

Большое количество несоответствий между разметками кореференции и ассоциативной анафоры у разных аннотаторов, вызванных неоднозначной интерпретацией данных отношений в тексте, естественным образом приводит к сомнениям в целесообразности осуществления такой разметки на большом корпусе текстов. Маловероятно (хотя мы этого не исключаем и продолжаем над этим работать), что разметка такого уровня сложности может быть в ближайшем будущем использована для успешного тестирования автоматически размеченных данных. Возможно, что исключение из разметки кореференции родовых понятий и девербативов сделает ее более точной, и соответствие между разметками

разных аннотаторов увеличится. Однако это не так просто сделать, так как границы между этими группами весьма размыты[8]. С другой стороны, есть все основания считать такую разметку кореферентной и анафорической информации очень ценной для лингвистических исследований. Анализ размеченного корпуса и сравнение параллельных разметок разных разметчиков дает возможность увидеть некоторые закономерности текста, незаметные при работе одного исследователя над небольшим объемом текстов. Множественность интерпретаций (даже очень простого) текста открывает и помогает решить также много вопросов психолингвистических исследований.

[8] См. Nedoluzhko 2007.

Литература

1. *Cohen J.* A coefficient of agreement for nominal scales. // Educational and Psychological Measurement, 20(1), 1960. С. 37–46.
2. *Chinchor N.* MUC-4 Evaluation Metrics // Proc. of the Fourth Message Understanding Conference, 1992. С. 22–29.
3. *Hajičová E.* и др. PDT 2.0 — Guide. UFAL & CKL, 2006. Доступно на http://ufal.mff.cuni.cz/pdt2.0/
4. *Hirschman L.* MUC-7 coreference task definition. Version 3.0. 1997.
5. *Krasavina O., Chiarcos Ch.* PoCoS — Potsdam Coreference Scheme. Proc. of ACL 2007, Prague, Czech Republic 2007.
6. *Kučová L.* и др. Anotování koreference v Pražském závislostním korpusu. ÚFAL/CKL Technical Report TR-2003-19. 2003.
7. *Nědolužko A.* Zpráva k anotování rozšířené textové koreference a bridging vztahů v Pražském závislostním korpusu. Technical report. Institute of formal and applied linguistics, Charles University, Prague. 2007. Доступно на http://ufal.mff.cuni.cz/~nedoluzko/koref_anot/manual_RK_kratky.pdf
8. *Nědolužko A., Mírovský J., Pajas P.* The Coding Scheme for Annotating Extended Nominal Coreference and Bridging Anaphora in the Prague Dependency Treebank. // Proceedings of ACL-IJCNLP 2009, Linguistic Annotation Workshop (LAW III). Suntec, Singapore, 2009.
9. *Poesio M.* The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. // Proceedings of SIGDIAL. Boston, 2004a.
10. *Poesio M., Delmonte R., Bristot A., Chiran L., Tonelli S.* The VENEX corpus of anaphora and deixis in spoken and written Italian. 2004б. Доступно на http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf
11. *Recasens M., Antònia Martí M.* AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources and Evaluation. 2010.
12. *Недолужко А., Гаич Я.* Синтаксически аннотированный корпус чешского языка. // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Выпуск 7 (14) 2008 С. 400–406.