

Гончаров А.А., Ярошенко П.В.

Имплицитные логико-семантические отношения: опыт анализа на материале параллельного корпуса

Аннотация. В статье описан подход к анализу имплицитных логико-семантических отношений. Материалом исследования послужили тексты двух параллельных подкорпусов Национального корпуса русского языка – английского и французского. Рассмотрен метод поиска примеров с имплицитными логико-семантическими отношениями, а также особенности разметки найденных примеров с использованием надкорпусной базы данных.

Ключевые слова. Параллельный корпус, база данных, имплицитность, логико-семантические отношения.

1. Введение

В настоящее время одним из актуальных направлений в лингвистике является анализ дискурса. В рамках этой проблематики исследуются, в частности, логико-семантические отношения (ЛСО) между единицами с пропозициональным содержанием. С точки зрения маркированности ЛСО могут быть эксплицитными – выраженными при помощи языкового средства (коннектора или иного показателя), и имплицитными – не выраженными ни одним из этих средств. Так, в примере (1) содержится эксплицитное ЛСО причины (выражено при помощи *так как*), а в примере (2) – имплицитное.

(1) *Они хотели поехать на эту конференцию, так как тематика их заинтересовала.*

(2) *Они хотели поехать на эту конференцию, тематика их заинтересовала.*

Настоящая статья посвящена имплицитным ЛСО и опыту их изучения на материале параллельного корпуса.

2. Поиск имплицитных ЛСО в параллельном корпусе

Следует отметить, что уже на этапе поиска примеров с имплицитными ЛСО для их последующего анализа могут возникать затруднения ввиду отсутствия показателя. По этой причине особенно актуальным представляется использование параллельного корпуса в качестве материала исследования: ЛСО, не имеющее показателя в тексте на одном языке, может быть эксплицитным в тексте на другом языке. В работе [Wu et al. 2016] этот феномен был назван «bilingual explicit / implicit mismatch». Таким образом, параллельные тексты способны обеспечить возможность целенаправленного поиска примеров, которые с высокой вероятностью содержат имплицитное ЛСО нужного типа (предполагается, что подобные различия в маркированности не оказывают негативного влияния на семантику текста в целом).

В качестве материала нашего исследования выступают тексты двух параллельных подкорпусов Национального корпуса русского языка – английского и французского. Для формирования выборки примеров с имплицитными ЛСО применяется метод поиска с исключением в параллельных текстах, представленный в работе [Гончаров 2023]. Суть метода состоит в том, чтобы получить список таких пар фрагментов параллельных текстов, где во фрагменте текста на одном языке показатели искомого ЛСО имеются, а во фрагменте текста на другом языке показатели этого же ЛСО, напротив, отсутствуют.

Например, чтобы в русско-английских параллельных текстах найти такие пары, где русскоязычный фрагмент с высокой вероятностью содержит имплицитное ЛСО причины, составлялся запрос, позволяющий получить список пар фрагментов, удовлетворяющих двум критериям: 1) во фрагменте оригинала отсутствуют лексические единицы, способные выражать ЛСО причины (в частности, *ибо, потому что, потому как, поскольку, так как, ведь, оттого что, из-за, за, от, ввиду, вследствие, затем что, тем более что, в связи с*); 2) во фрагменте перевода имеется единица *because*. При составлении запроса приходилось, с одной стороны, с осторожностью подходить к составлению списка исключаемых единиц – если единица многозначна и при этом частотна, можно потерять ценные примеры, где искомое ЛСО имплицитно, а эта единица имеется, но используется в другом значении, – а с другой стороны, стремиться к сокращению «шума» в выдаче. Прежде чем добавить единицу в список исключаемых, всегда производился анализ корпусных данных.

Метод позволяет значительно ускорить формирование набора размеченных вручную примеров имплицитных ЛСО (не прибегая к синтетическим данным): единственной известной нам альтернативой был бы сплошной просмотр текстов.

С использованием вышеупомянутого метода в рамках исследования была создана выборка, включающая 1002 переводных соответствия с имплицитными ЛСО в русском языке, где в другом языке имеется показатель ЛСО. Выборка включает ЛСО причины и противопоставления (направления перевода: русский–французский, русский–английский, французский–русский, английский–русский).

3. Аннотирование имплицитных ЛСО с использованием надкорпусной базы данных

На следующем этапе было произведено первичное аннотирование созданной выборки переводных соответствий с имплицитными ЛСО. Результатом этой работы являются аннотированные переводные соответствия (АПС), которые сохраняются в базе данных, представляющей собой своего рода надстройку над корпусом – надкорпусной базе данных [Кружков 2021].

Структура АПС детально представлена в работе [Гончаров, Ярошенко 2024]). Здесь лишь перечислим основные элементы АПС: 1) блоки текста оригинала и перевода (объем блоков текста должен быть необходимым и достаточным для анализа ЛСО); 2) наборы признаков блоков текста (делятся на группы в зависимости от того, к какой части блока текста они относятся); 3) набор признаков переводного соответствия (необходимо для случаев, когда тексты оригинала и перевода имеют значимые различия). В процессе аннотирования внутри каждого блока текста выделялись аргументы анализируемого ЛСО – отрезки текста, связанные при помощи этого ЛСО – и им присваивались метки «аргумент 1» и «аргумент 2» соответственно. Выбор нужной метки, с учетом опыта [Bunt, Prasad 2016], для асимметричных отношений осуществляется исходя из семантики аргумента (т.е. линейный порядок не учитывается), а для симметричных – напротив, исключительно из линейного порядка аргументов. Например, для ЛСО причины (асимметричное ЛСО) меткой «аргумент 1» всегда обозначается тот, что содержит причину того, что описано во втором аргументе, а для ЛСО противопоставления (симметричное ЛСО) – меткой «аргумент 1» обозначается тот, что находится в тексте линейно раньше, чем

второй аргумент отношения. Следует отметить, что такое разграничение представляет собой некоторое упрощение, так как примеры из текстов на естественном языке показывают, что даже те отношения, которые считаются симметричными, не всегда допускают возможность перестановки аргументов без изменения смысла.

Примеры АПС приведены на рис. 1 и 2. Рисунки представляют собой скриншоты из приложения для работы с базой данных, используемой в исследовании. Аргументы ЛСО выделены цветом: зеленым – аргумент 1, оранжевым – аргумент 2.

АПС № 364		Источник
Башкирец с трудом шагнул через порог (он был в колодке) и, сняв высокую свою шапку, остановился у дверей [...].	Le Bachkir franchit le seuil avec peine, car il avait aux pieds des entraves en bois. Il ôta son haut bonnet et s'arrêta près de la porte [...].	[А. С. Пушкин. Капитанская дочка (1836) La fille du capitaine (пер. L. Viardot; 1853)]
NoMarker	car	Признаки переводного соответствия
Признаки блока текста в целом		
ЛСО: <причина (пр.)>	ЛСО: <причина (пр.)>	
Признаки аргумента 1		
Синтаксическая характеристика arg1: <вставная конструкция (1)>	отсутствуют	
Признаки аргумента 2		
Субъект-оцен. лексика в arg2: <оценка (2)>	Субъект-оцен. лексика в arg2: <оценка (2)>	
с, трудом	avec, peine	
Признаки аргументов 1 и 2		
Соотношения на уровне лексики: <анафора>	Соотношения на уровне лексики: <анафора>	
Башкирец, он	Bachkir, il	

Рис. 1. Пример АПС с имплицитным ЛСО причины

АПС № 1366		Источник [С. Д. Довлатов. Заповедник (1983) Domaine Pouchkine (пер. Chr. Zeytounian- Belous, 2004)] Признаки переводного соответствия <i>отсутствуют</i>
— Я Михал Иванович не знаю. Он производит хорошее впечатление.	— Je ne connais pas Mikhail Ivanytch, mais il m'a fait bonne impression.	
NoMarker	mais	
Признаки блока текста в целом		
ЛСО: <противопост.>	ЛСО: <противопост.>	
Признаки аргумента 1		
Предикат в arg1: <отрицание (1)> ■ не	Предикат в arg1: <отрицание (1)> ■ ne, pas	
Признаки аргумента 2		
Субъект-оцен. лексика в arg2: <оценка (2)> ■ хорошее	Субъект-оцен. лексика в arg2: <оценка (2)> ■ bonne	
Признаки аргументов 1 и 2		
Пунктуация и расположение в тексте: <разные предложения>	Соотношения на уровне лексики: <анафора> ■ Mikhail, Ivanytch, il	
Соотношения на уровне лексики: <анафора> ■ Михал, Иванович, Он		

Рис. 2. Пример АПС с имплицитным ЛСО
противопоставления

Отдельной сложной задачей было составить перечень признаков для разметки АПС. При его составлении мы ориентировались (1) на анализ корпусных данных, (2) на работы по лингвистической семантике, а также (3) учитывали работы, касающиеся проблематики автоматического распознавания имплицитных ЛСО (implicit discourse relation recognition). В этих работах исследователи, среди прочего, занимались т.н. «конструированием признаков» (feature engineering) [Das, Taboada 2018; Xiang, Wang 2023], но используемые ими признаки, как правило, не предполагали никакого семантического анализа, будучи формальными (напр., число слов в клаузе, число именных групп, число местоимений и т.п.).

В результате был составлен предварительный перечень признаков, которые можно классифицировать на основании того, к какой именно части блока текста они относятся. В соответствии с этой логикой, перечень признаков был разделен на три класса: (1) признаки блока текста в целом (тип ЛСО); (2) признаки,

относящиеся к одному из аргументов (синтаксическая характеристика; указание на лексические единицы, значимые для анализа ЛСО); признаки, относящиеся к обоим аргументам (типы соотношения аргументов на лексическом уровне; на уровне пунктуации); признаки переводного соответствия (например, можно указать на то, что в переводе изменена структура блока текста оригинала).

Стало ясно, что причины выбора разметчиком того или иного признака могут быть не вполне очевидными. Чтобы, с одной стороны, эксплицировать эти причины, и с другой стороны, упростить последующую работу с массивом АПС, было решено разделить все признаки на две категории: способные и неспособные иметь лексический показатель. Если признак не может иметь лексический показатель, для разметчика ничего не меняется; в противном случае разметчик должен указать, какое именно слово в блоке текста заставило его выбрать заданный признак (примеры см. на рис. 1 и 2, где для признаков «Оценка», «Отрицание» и «Анафора» выбраны лексические показатели, что отмечено флажками). Такое решение одновременно и усложнило процедуру аннотирования, и упростило ее: с одной стороны, добавился дополнительный этап, с другой – инструкция стала намного более четкой. Главным же преимуществом предлагаемой схемы аннотирования стало получение данных, размеченных значительно более последовательным образом.

4. Заключение

В статье показано, что в рамках исследования имплицитных ЛСО параллельный корпус имеет ряд преимуществ перед монологическим материалом. Во-первых, он обеспечивает возможность целенаправленного поиска примеров таких

отношений. Во-вторых, сопоставительный анализ фрагментов, где в одном языке ЛСО выражено специализированным показателем, а в другом – имплицитно, позволяет опираться не только на интроспекцию, но и на решения, принятые переводчиком, который рассматривается как независимый эксперт. В-третьих, факт выбора различных языковых средств для выражения аналогичного содержания может показать исследователю, какой именно вклад в семантику фрагмента текста вносит каждое средство в отдельности. Разработанные информационные ресурсы (база данных и приложение для работы с ней) и методика аннотирования призваны стать основой для создания набора данных, необходимого для изучения имплицитных ЛСО на русскоязычном материале, а также обеспечить возможность дальнейшего теоретического осмысления природы этих отношений.

Литература

1. Гончаров А.А. (2023), Поиск с исключением в параллельных текстах. *Системы и средства информатики*, 33 (4), с. 102–114.
2. Гончаров А.А., Ярошенко П.В. (2024), Принципы аннотирования имплицитных логико-семантических отношений в параллельных текстах. *Информатика и ее применения*, 18 (3), с. 106–114.
3. Круглов М.Г. (2021), Концепция построения надкорпусных баз данных. *Системы и средства информатики*, 31 (3), с. 101–112.
4. Bunt H., Prasad R. (2016), ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. *12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation: Proceedings*, Portorož, Slovenia, pp. 45–54.
5. Das D., Taboada M. (2018), RST Signalling Corpus: a corpus of signals of coherence relations. *Language Resources and Evaluation*, 52, pp. 149–184.

6. Wu Ch., Shi X., Cheng Y., Huang Y., Su J. (2016) Bilingually-constrained Synthetic Data for Implicit Discourse Relation Recognition. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, ACL, pp. 2306–2312.

7. Xiang W., Wang B. (2023), A survey of implicit discourse relation recognition. *ACM Computing Surveys*, 55 (12), 34 p.

References

1. Goncharov A.A. (2023), Poisk s isklyucheniem v parallel'nykh tekstakh [Search with Exclusion in Parallel Texts]. *Sistemy i Sredstva Informatiki* [Systems and Means of Informatics], 33 (4), pp. 102–114. (In Russ.)

2. Goncharov A.A., Iaroshenko P.V. (2024), Printsipy annotirovaniya implitsitnykh logiko-semanticheskikh otnosheniy v parallel'nykh tekstakh [Implicit logical-semantic relations in parallel texts: annotation principles]. *Informatika i ee Primeneniya* [Informatics and Applications], 18 (3), pp. 106–114. (In Russ.)

3. Kruzhkov M.G. (2021), Kontseptsiya postroeniya nadkorpornykh baz dannykh [Conceptual framework for supracorpora databases]. *Sistemy i Sredstva Informatiki* [Systems and Means of Informatics], 31 (3), pp. 101–112. (In Russ.)

4. Bunt H., Prasad R. (2016), ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. *12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation: Proceedings*, Portorož, Slovenia, pp. 45–54.

5. Das D., Taboada M. (2018), RST Signalling Corpus: a corpus of signals of coherence relations. *Language Resources and Evaluation*, 52, pp. 149–184.

6. Wu Ch., Shi X., Cheng Y., Huang Y., Su J. (2016) Bilingually-constrained Synthetic Data for Implicit Discourse Relation Recognition. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, ACL, pp. 2306–2312.

7. Xiang W., Wang B. (2023), A survey of implicit discourse relation recognition. *ACM Computing Surveys*, 55 (12), 34 p.

Analysis of Implicit Discourse Relations on the Material of Parallel Corpus

Abstract. The article focuses on an approach to the analysis of implicit discourse relations. The material of the study includes texts of two parallel subcorpora of the Russian National Corpus – English and French. The paper discusses the method for retrieval of implicit discourse relations instances in parallel texts. Applying the aforementioned method, we created a sample of 1 002 translation correspondences with implicit discourse relations in Russian, where a text span in English or French contains a relation marker. The sample includes fragments with discourse relations of cause and contrast. The specifics of translation correspondences annotating using the supracorpora database are considered. Examples of annotated translation correspondences are given.

Keywords. Parallel corpus, database, implicitness, discourse relations.