УДК 81'322.4
DOI https://doi.org/10.32841/2409-1154.2019.41.1.28

Нагиева Парвин,
докторант
Бакинского государственного университета

ТРУДНОСТИ ПЕРЕДАЧИ ПОДЧИНИТЕЛЬНЫХ СОЮЗОВ В МАШИННОМ ПЕРЕВОДЕ

Анотація. Статтю присвячено проблемі уявлення семантики складнопідрядних пропозицій в мовно-незалежній формі, яка потім може бути використана для машинної обробки інформації. Розуміння складнопідрядних речень може викликати суттєві труднощі у людини, для систем штучного інтелекту (ІІ) рішення зазначеного завдання протягом тривалого часу і зовсім уявлялося неможливим. Системи і алгоритми обробки і інтерпретації повідомлень стикаються з істотними труднощами і невдачами через багатозначності підрядних спілок, їх високу модальність, взаємозв'язок із соціально-природним навколишнім середовищем і когнітивними особливостями пізнання особистістю зазначеного середовища. У статті розглядаються можливі шляхи вирішення цієї проблеми.

Нині переклад відносно успішно вирішується методами глибокого навчання (deep learning). На відміну від перекладів, подання інформації в мовно-незалежній формі вимагає більшої точності і стандартизації, позаяк завдання глибокого навчання в цьому випадку полягає у перекладі вихідного тексту ПМ в спеціальну мовно-незалежну конструкцію. Ці конструкції повинні включати в себе все різноманіття використання спілок у природних мовах. Один із перших алгоритмів SYSCONJ для обробки синтаксичних конструкцій у системі LUNAR вже розроблений.

Логіка проведення дослідження вимагає, перш за все, виявлення сутності підрядних союзів. Під союзом у вітчизняній науковій літературі можна розуміти «розряд слів, що показує зв'язок між словами, складниками мова». Звернення до логічної модальності спілок, мовно-незалежне їх подання, а також використання відповідних позначок дозволяє вирішити проблему уявлення спілок ПМ, зробити їх «зрозумілими», які «читаються» системами ІІ, а також передати логічні вирази ПМ лінгвістичної форми.

Ключові слова: складнопідрядні союзи, складнопідрядні речення, подання знань у базах даних, машинний переклад.

Введение. Вопросы взаимосвязи между естественными и искусственными языками, обработки, интерпретации и языково-независимого представления ЕЯ в системах ИЯ имеют большое теоретическое и прикладное значение. Автоматическая обработка текстов ЕЯ используется для машинного перевода, информационного поиска, поддержки диалога на ЕЯ (чат-боты), автоматизации подготовки и редактирования текстов ЕЯ, обучении ЕЯ. Адекватная интерпретация текстов ЕЯ является ключевой задачей при разработке систем искусственного интеллекта (ИИ). Проблема адекватного представления текстов ЕЯ средствами ИЯ не нова. Первые попытки использования лингвистики компьютерными науками были предприняты еще в середине прошлого столетия. Нередко пессимистически оценивалась даже возможность точного перевода научного текста несмотря на его очень скромный тезаурус, вплоть до понимания того, что «машинный перевод научного текста общего характера не осуществлен и не будет осуществлен в ближайшей перспективе» [8, с. 61].

Трудности у систем ИИ возникают на каждом этапе обработки текста: лексика, синтаксис, семантика, прагматика. Поэтому системы ИИ работают поэтапно: сначала выявляется семантика, заключенная в отдельно взятых лексических единицах (клаусах), а затем уже происходит интегрирование клаусов через их логическую связь для выявления семантики, заключенной в предложении в целом. Естественные языки логически соединяют клаусы с помощью союзов, которые в простейших случаях сочинительных союзов являются аналогами булевских логических операндов в машинных языках (например, «и», «или», «нет», «поэтому»). Но естественный язык выходит далеко за пределы булевой алгебры, он выступает не просто средством передачи «голой» информации, но и служит инструментом передачи эмоциональных (модальных) состояний и оценок явлений, свойственных человеку, но абсолютно не свойственных ИИ (по крайней мере, на современном этапе развития) [18, с. 2]. ЕЯ использует различные средства для передачи модальности. Основная нагрузка для передачи модальности в ЕЯ ложится на союзы, особенно на подчинительные.

В отличие от сочинительных союзов, подчинительные выступают средством выражения. Соответственно, ограниченные ресурсы систем ИИ сталкиваются с необходимостью передачи безграничности человеческих отношений к действительности, что обусловливает актуальность исследования.

Целью исследования выступает выявление механизмов передачи подчинительных союзов в языко-независимой форме.

Извлечения информации, базы данных. В основе всех алгоритмов, направленных на извлечение информации, лежит совокупность методов обработки текстов ЕЯ, которая может быть условно сведена к статистическому и лингвистическому подходам [2, с. 30].

В основе *статистического подхода* лежит предположение о том, что содержание текста может быть отражено на основании выявления наиболее часто встречающихся слов. Суть статистического подхода заключается в подсчёте количества вхождений лексемы в исследуемый текст (корпус текстов). Одним из эффективных подходов, основанных на статистическом анализе, является латентно-семантическое индексирование. Латентно-семантический анализ представляет собой «теорию и метод контекстнозависимых значений слов при помощи статистической обработки больших наборов текстовых данных» [2, с. 30]. В основу метода заложена идея, что совокупность всех контекстов, в которых встречается либо не встречается лексема, задает множество обоюдных ограничений, которые

в значительной степени позволят определить смысловые значения лексических единиц множества лексем между собой.

Автоматизированные переводы с ЕЯ на ЕЯ. Следует отметить, что передача сложноподчиненных конструкций представляет собой одну из наиболее проблематичных областей автоматической обработки любого языка. Первые модели разбора сложных предложений английского языка были предприняты сразу же после периода «столкновения с реальностью» (1966 – 1973 гг.), когда ученые пришли к выводу о невозможности достижения адекватности в процессе машинного перевода [8, с. 62]. Сегодня перевод относительно успешно решается методами глубокого обучения (deep learning). В отличии от переводов, представление информации в языково-независимой форме требует большей точности и большей стандартизации, так как задача глубокого обучения в этом случае заключается в переводе исходного текста ЕЯ в специальную языково-независимую конструкцию. И эти конструкции должны включать в себя все многообразие использования союзов в естественных языках.

Один из первых алгоритмов SYSCONJ для обработки синтаксических конструкций в речевой вопросно-ответной системе LUNAR был разработан В. Вудсом в 1973 г. [22]. Разработанный алгоритм успешно подвергал парсингу короткие конструкции (например, *John drove his car through and completely demolished a plate glass window*), однако имел существенные недостатки [19, с. 81].

1) Использование этого алгоритма для обработки различных типов было слишком дорогостоящим и малоэффективным;

2) алгоритм имел низкую степень детерминированности, соответственно, легко приводил к существенным искажениям смысла.

В 1981 г. был представлен алгоритм Блэквэлла WRD AND arc [16]. Указанный алгоритм позволял осуществлять парсинг отдельных сложных конструкций, однако процесс был слишком громоздким, а результаты – зачастую недостоверными.

Система машинного перевода, разработанная в 1982 г. под руководством М. Нагао [20], успешно обрабатывала синтаксические конструкции, состоящие из двух рядом стоящих слов, однако конструкции типа “Noun + Preposition + Noun”, “Adjective + Noun + Conjunction + Noun” оставались недоступны для анализа.

На сегодняшний день большинство систем машинного перевода основывается на модели «последовательность-в-последовательность» (seq2seq) нейронного машинного перевода (НМП). Как отмечает С.М. Калинин, НМП – «это относительно новый подход к решению проблемы машинного перевода, получивший широкое распространение в последние годы» [3, с. 11]. Функционирование подхода основывается на использовании нейронных сетей, которые по своей структуре напоминают строение человеческого мозга. Главным преимуществом подобных систем выступает возможность самообучения.

Современные модели seq2seq имеют во многом аналогичную структуру, включают кодер, декодер и механизм отслеживания (attention mechanism) [3, с. 11]. Функционирование кодера основывается на обработке входящих символов изначально в скрытые, а затем выходные нейроны. В графической форме указанный процесс может быть представлен в следующем виде:

(1) $x = (x_1 \ldots x_j) \rightarrow h = (h_1 \ldots h_j) \rightarrow y = (y_1 \ldots y_j)$, где x – входные символы, h – скрытые символы, y – выходные сигналы

На начальном этапе осуществляется парсинг сложного предложения, который представляет собой перевод текста ЕЯ в набор меток. Он может быть представлен следующим образом [15]:

1) синтаксические характеристики: например, *S (sentence) – начальный символ; MC – Main Clause, SP – Subordinate Clause* и прочих;

2) лексико-семантические характеристики;

3) морфологические характеристики.

Перевод сложноподчиненного предложения на английский язык может быть представлен следующим образом:

I cannot take the initiative ***if*** *I do not respect the boss (Google Translator)* – как отражает приведенный перевод, отмечаются некоторые семантические деформации в процессе перевода существительного «начальник», которое переведено как “boss”. Если лексема «начальник» в русском языке обозначает «должностное лицо, руководящее, заведующее чем-нибудь» [13], то босс – это жаргонная номинация начальника. В этом случае отмечается передача нейтральной литературной лексемы при помощи жаргонизма. Иначе говоря, несмотря на то, что системе удалось передать прагматическую сущность синтаксической конструкции, передать значение подчинительного союза, отношения подчиненности, представленные в сообщении оригинала, семантические и прагматические аспекты, модальность передачи оригинального сообщения зачастую остаются недоступными.

Представление логических конструкций сложно подчиненных союзов в языково-независимой форме. Важнейшим звеном в технологической цели компьютерной обработки ЕЯ является формально-логическое представление его конструкций: «Высказывание на естественном языке, будучи переведенным в формально-логическое представление, уже может быть «понятным» компьютеру и обработано в соответствии с конкретными задачами пользователя» [7, с. 83].

Основные трудности с обработкой и интерпретацией конструкций с подчинительными союзами обусловлены особенностями самих подчинительных союзов, включая следующие [21]:

1. Трудности с анализом: подчинительные союзы любого языка имеют множество значений, реализуют ряд функций, объединяя главное и подчиненное предложениями, сливаясь по форме и значению с сочинительными союзами и другими частями речи, воспроизводя бесконечное множество синтаксических конструкций и семантических образований.

2. Многозначность союзов.

3. Эллипсисы: синтаксические конструкции могут включать значение союза имплицитно.

Классификации союзов в лингвистике. Логика проведения исследования требует прежде всего выявления сущности подчинительных союзов. Под союзом в отечественной научной литературе может пониматься «разряд слов, показывающий связь между словами, составляющими речь» [4, с. 65]; служебная часть речи, «которая включает в себя слова, соединяющие или разъединяющие слова, словосочетания или предложения со стороны тех или иных отношений» [15, с. 505]; «служебные слова, выражающие смысловые отношения между однородными членами простого предло-

жения и между частями сложного предложения – сложносочиненного и сложноподчиненного» [1, с. 268]; «служебные слова, выражающие синтаксические отношения между членами предложения, частями сложного предложения и отдельными предложениями» [10, с. 248].

Как отражают приведенные дефиниции, большая часть исследователей подчеркивает тот факт, что союзы реализуют связующую роль между членами словосочетания, частями сложного предложения и предложениями в целом, внося логико-грамматический акцент в подчинительную или сочинительную связь сложного предложения. По своей сути, союзы представляют собой грамматическое средство взаимосвязи реалий / процессов социально-природной окружающей среды и результатами закрепления познания окружающей социальной среды в языковой системе. Как отмечает М.В. Ляпон, «уже сам по себе выбор связующего средства, с помощью которого инициатор сообщения соединяет фрагменты информации, когда он строит высказывание в форме сложного предложения, есть не что иное, как операция умозаключения, поскольку этот выбор предопределен тем выводом, к которому говорящий приходит, оценивая и квалифицируя отношения между соединяемыми фрагментами, то есть подвергая информацию специальной логической обработке» [5, с. 9]. Иначе говоря, союз представляет собой грамматический индикатор реализации когнитивных процессов личности в процессе познания социально-природной среды и выведения умозаключений, закрепленных в языковой системе. Другими словами, союзы, например русского языка, обладают высокой модальностью – «функционально-семантической категорией, которая имеет отношение к действительности и к мнению говорящего» [14, с. 35].

Лингвистика союзов и формальная логика их машинного представления. Принимая во внимание выявленные лингвистические значения союзов, представляется важным перевод их в логическую форму машинного представления. Важно добавить, что «лингвистическая модальность не только не исключает логическую модальность, но и базируется на ней» [9, с. 6]. Соответственно, большую часть значений мы можем свести к значению логических союзов, выражающих следование, конъюнкцию, дизъюнкцию, эквивалентность, импликацию, отрицание [7, с. 86]. В частности, союз ***и*** выражает перечисление слов, находящихся в однородных отношениях и обозначающих различные предметы, признаки, явления, образовывает *чистую конъюнкцию (и-отношения):*

И *жить хочу,* ***и*** *пить,* ***и*** *есть,*

Хочу тепла и света...(Твардовский).

Это же предложение, записанное в языково-независимой форме по отношению к союзам, имеет вид:

(Я хочу) <- *(*жить, пить, есть, тепло, свет)

В приводимых примерах клаусы помещаются в круглые скобки. Стрелка "**<-**" направлена на тот клаус, свойства которого описываются в предложении. Исходный союз заменяется на его языково-независимую форму. Союз **[и]** здесь и далее обозначается запятой.

При повторном союзе ***и – и,*** кроме перечисления, выражается усиление.

И *пращ,* ***и*** *стрела,* ***и*** *лукавый кинжал*

Щадят победителя годы (Пушкин).

(Щадят победителя годы) <- **[УСИЛЕНИЕ]** (пращ, стрела, *(*лукавый кинжал).

Союз ***ни*** (повторный) выражает такое же перечисление в отрицательных предложениях (с усилением) – противительная конъюнкция (ни-отношения, но-отношения, да-отношения, а-отношения):

Потом увидел ясно он,

Что и в деревне скука та же,

Хоть нет ***ни*** *улиц,* ***ни*** *дворцов,*

Ни *карт,* ***ни*** *балов,* ***ни*** *стихов* (Пушкин).

(Потом увидел ясно он, что и в деревне скука та же) <- **[УСИЛЕНИЕ]** *(***[НЕТ***]* *(улиц, дворцов, карт, балов, стихов).* Здесь отрицание ***[НЕТ]*** указывает на отсутствие перечисляемых далее объектов.

Для каждого объекта реального мира человек помнит большое количество его возможных свойств. Чтобы из всего множества свойств объекта выделить те, которые особенно важны в разговоре именно в этой ситуации (контексте), ЕЯ имеет специальные лингвистические средства, каждый ЕЯ свои. Когда мы хотим представить исходное предложение в языково-независимой форме, мы используем специальную метку **[УСИЛЕНИЕ].**

Люблю отчизну я, но странною любовью!

Не победит её рассудок мой,

Ни *слава, купленная кровью,*

Ни *полный гордого доверия покой,*

Ни *тёмной старины заветные преданья*

Не шевелят во мне отрадного мечтанья (Лермонтов).

(Люблю отчизну я, но странною любовью! Не победит её рассудок мой) <- ([**НЕТ** (шевелят во мне отрадного мечтанья) <- **[УСИЛЕНИЕ]** *(*слава, купленная кровью), *(*полный гордого доверия покой), *(*тёмной старины заветные преданья). Здесь отрицание [**НЕТ**] применяется к следующему за нею клаусу (шевелят во мне отрадного мечтанья). Следующий кластер перечисляет акцентированные признаки «отрадного мечтанья».

Как отражают приведенные примеры, обращение к логической модальности союзов, языково-независимое их представление, а также использование соответствующих меток позволяет решить проблему представления союзов ЕЯ, сделать их «понятными», «читаемыми» системами ИИ, а также передать логические отношения, выражаемыми в ЕЯ в лингвистической форме.

Статистика употребления союзов и этапы разработки языково-независимых логических конструкций. В соответствии с данными Национального корпуса русского языка, численность союзов составляет 7,9% от общего числа употребляемых частей речи (для сравнения: численность потребляемых числительных составляет лишь 1,7%) [6], что отражает высокочастотное использование служебной речи при оформлении высказываний ЕЯ. В сложившейся ситуации представляется актуальным выделение следующих этапов разработки языково-независимых логических конструкций:

1) принимая во внимание тот факт, что союзы ЕЯ представляют собой средство выражения лингвистической модальности, представляется актуальным, прежде всего, выявить сущность указанной категории, выявить совокупность модальных отношений, выражений в ЕЯ;

2) выявление модальных отношений, категорий позволит выявить совокупность союзов как средств отражения того или иного модального отношения личности, что позволит сгруппировать союзы ЕЯ в соответствии с выражаемыми видами модальности;

3) на третьем этапе работы представляется важным провести соответствие между лингвистической и логической модальностями, приведение указанных категорий в соответствие, единообразие;

4) разработка системы меток для языково-независимого представления с целью репрезентации отдельных лингвистических модальностей в структуре логических с целью конкретизации, дополнения логических модальностей отдельными аспектами лингвистических, выраженных в ЕЯ.

Литература:

1. Валгина Н.С., Розенталь Д.Э., Фомина М.И. Современный русский язык. Учебник. 6-е изд., перераб. и доп. М. : Логос, 2002. 528 с.
2. Диковицкий В.В., Шишаев М.Г. Обработка текстов естественного языка в моделях поисковых систем // Труды Кольского научного центра РАН. 2010. Вып. 1. С. 29–34.
3. Калинин С.М. Обзор современных подходов к улучшению точности нейронного машинного перевода // Rhema. Рема. 2017. № 2. С. 70–79.
4. Курбанова Р.Г. К характеристике союзов и союзных слов в русском и табасаранском языках // Известия Дагестанского государственного педагогического университета. Общественные и гуманитарные науки. 2013. № 4. С. 65–70.
5. Ляпон М. Смысловая структура сложного предложения и текст. Монография. М. : Наука, 1986. 200 с.
6. Национальный корпус русского языка. Электронный ресурс. Адрес доступа: http://www.ruscorpora.ru/corpora-stat.html.
7. Онтологические методы и средства обработки предметных знаний : монография / А.В. Палагин, С.Л. Крывый, Н.Г. Петренко. Луганск : Изд-во ВНУ им. В. Даля, 2012. 324 с.
8. Рассел С., Норвиг П. Искусственный интеллект. Современный подход. 2-е изд. М. : Вильямс, 2007. 1410 с.
9. Романова Т.В. Модальность. Оценка. Эмоциональность : монография. Нижний Новгород : НГЛУ им. Н.А. Добролюбова, 2008. 309 с.
10. Современный русский язык. Часть 2. Шанский Н.М., Тихонов А.Н. Словообразование. Морфология. В 3 частях. 2-е изд., испр. и доп. М. : Просвещение, 1987. 256 с.
11. Солганик Г.Я. О текстовой модальности как семантической основе текста // Структура и семантика художественного текста: Доклады VII Международной конференции МГОПУ. М., 1999. С. 364–372.
12. Толковый словарь Ушакова. Д.Н. Ушаков. 1935–1940. Электронный ресурс. Адрес доступа: https://dic.academic.ru/dic.nsf/ushakov/878946.
13. Шакирзянова Р.М. Роль модальности в лингвистике // Ученые записки Казанской государственной академии ветеринарной медицины им. Н.Э. Баумана. 2011. № 1–2. С. 32–35.
14. Шахматов А.А Синтаксис русского языка. М. : Эдиториал УРСС, 2001. 624 с.
15. Blackwell S.A. (1981) "*Processing Conjunctions in an ATN Parser"*. Unpublished M. Phil. Dissertatation, University of Cambridge.
16. Dong, Ngan and Nguyen, Kim Anh (2018) *Attentive Neural Network for Named Entity Recognition in Vietnamese* CoRRabs/1810.13097 Electronic Resource. Access mode: https://arxiv.org/pdf/1810.13097.pdf.
17. Huang X. (1983). Dealing with conjunctions in a machine translation environment. In *Proceedings of European A CL Conference*: pp. 81– 85.
18. Nagao M., Tsijii J., Yada K. and Kakimoto T. (1982) An English Japanese Machine Translation System of the Titles of Scientific and Engineering Papers. In Horecky J. (ed), *COLING 82,* NorthHolland Publishing Company.
19. Okumura, Akitoshi and Muraki, Kazunori (1994) *Symmetric Pattern Matching Analysis for English Coordinate Structures*. Electronic Resource. Access mode: http://mt-archive.info/ANLP-1994-Okumura.pdf.
20. Woods W. (1973) "*An Experimental Parsing System for Transition Network Grammar". In Rustin, R. (ed), Natural Language Processing*, Algorithmic Press, N.Y.

Nagiyeva P. The difficulties of translation of subordinate conjunctions in machine translation

Summary. The article is devoted to the problem of representing the semantics of complex sentences in a language-independent form, which can then be used for machine information processing. Understanding complex sentences can cause significant difficulties even for a person; for artificial intelligence (AI) systems, the solution of this problem for a long time seemed completely impossible. Systems and algorithms for processing and interpreting messages encounter significant difficulties and failures due to the ambiguity of subordinate unions, their high modality, their relationship with the social and natural environment, and the cognitive characteristics of cognition by the personality of this environment. The article discusses possible solutions to this problem. Today, translation is relatively successfully solved by deep learning methods. Unlike translations, the presentation of information in a language-independent form requires greater accuracy and greater standardization, since the task of deep learning in this case is to translate the source text of the NJ into a special language-independent design. And these constructions should include all the diversity of the use of unions in natural languages. One of the first SYSCONJ algorithms for processing syntactic constructions in the LUNAR speech question-answer system was developed.

The logic of this study requires, first of all, identifying the essence of subordinate unions. Under the union in the domestic scientific literature can be understood as "a category of words showing the relationship between the words that make up speech". The appeal to the logical modality of unions, their language-independent representation, as well as the use of appropriate labels, allows us to solve the problem of representing the unions of NL, make them "understandable", "readable" by AI systems, and also convey logical relationships expressed in NL in linguistic form.

Key words: complex subject unions, complex subject sentences, representations of knowledge in databases, machine translation.