Вісник Національного технічного університету "ХПІ", 2021, № 1 (5)
ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

УДК 510.635:004.891(045) DOI: 10.20998/2411-0558.2021.01.06

А. І. ВАВІЛЕНКОВА, д-р техн. наук, проф., НАУ, Київ

ВЗАЄМОЗВ'ЯЗОК ТИПІВ ОКРЕМИХ ФОРМ ЛОГІКО-ЛІНГВІСТИЧНИХ МОДЕЛЕЙ ТА ВИДІВ РЕЧЕНЬ ПРИРОДНОЇ МОВИ

Запропоновано апарат логіко-лінгвістичного моделювання для вилучення знань з електронних текстових документів. Виокремлено різні типи форм логіко-лінгвістичних моделей для формалізованого представлення складних речень природної мови. Сформульовано правила та здійснено класифікацію окремих форм логіко-лінгвістичних моделей залежно від синтаксичних конструкцій речень природної мови. Іл.: 1. Бібліогр.: 12 назв.

Ключові слова: логіко-лінгвістична модель; вилучення знань; окрема форма; речення природної мови; електронний текстовий документ.

Постановка проблеми. Сьогодні актуальною для сфери інформаційних технологій задачею залишається вилучення знань із електронних текстових документів з подальшою можливістю їх порівняння, пошуку та створення нового контенту. Інструментом для розв'язання цієї задачі може служити логіко-лінгвістичне моделювання. Зокрема, для формалізованого представлення простих речень природної мови автором було розроблено шаблон [1], логіко-лінгвістичну модель, що складається з простих предикатів першого порядку і до складу якої входять відношення, суб'єкт, об'єкт, предмет відношення та їх характеристики відповідно. Логіко-лінгвістична модель для кожного речення природної мови набуває окрему форму залежно від того, скільки граматичних основ у реченні, чи присутні однорідні члени, а також, які логічні та контекстуальні зв'язки між простими реченнями у межах складного. Якщо два перших фактори описуються у вигляді різних типів окремих форм логіко-лінгвістичних моделей, то такий фактор, як логічні та контекстуальні зв'язки потребує ретельного дослідження, чому і присвячено матеріал даної статті.

Аналіз літератури. Проблемою формального представлення та вилучення знань з текстової інформації займається багато вітчизняних [2, 3] та зарубіжних [4, 5] вчених базуючи свої дослідження на різних типах моделей представлення знань. Так, у статті австрійських авторів [6] запропоновано алгоритм вилучення інформації з веб-таблиць, що базується на дослідженні топології таблиць та розрахунку відстані між колонками. Популярною тенденцією серед іноземних розробників у

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

сфері Natural Language Processing є використання методу N-грам [7].

В області логіко-лінгвістичного моделювання, що поєднує формальний математичний апарат та основи комп'ютерної лінгвістики, також зроблено спроби застосувати логіко-лінгвістичні моделі для вилучення фактів [8], а також інтерпретації предикатів як форм опису граматичних характеристик розміщення фактів у реченні природної мови [9]. Розглянуті матеріали дозволяють реалізувати процес формалізації та вилучення знань з простих речень природної мови з певною імовірністю, проте при аналізі складних речень природної мови, а тим більше цілих електронних текстових документів, імовірність коректної обробки текстової інформації зменшується.

Мета статті – здійснити класифікацію окремих форм логіко-лінгвістичних моделей представлення знань залежно від типів речень природної мови, які вони описують, що в подальшому дасть можливість розробити алгоритм відновлення текстової інформації з формальних моделей.

Логіко-лінгвістичні моделі складних речень природної мови. Залежно від типу речення природної мови, яке підлягає формалізації, логіко-лінгвістична модель [10] може приймати вигляд однієї із своїх окремих форм, класифікацію яких продемонстровано на рис. 1.

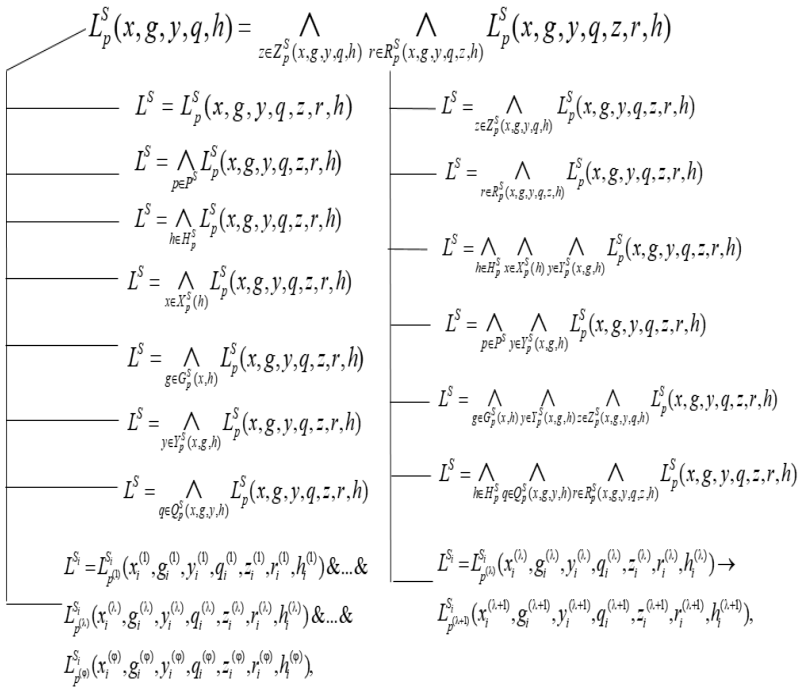

Рис. 1. Класифікація окремих форм логіко-лінгвістичних моделей

Складне речення – це речення, в якому одна складова є незалежна, а друга – підпорядкована, що приєднується сполучником підрядності або за допомогою сполучного слова. Таке речення можна інтерпретувати за допомогою декількох простих предикатів [11], об'єднаних певними логічними операціями. За синтаксичним відношенням складнопідрядні речення поділяються на означальні, з'ясувальні та обставинні.

Саме від типу складного речення і залежить, яка логічна операція буде використана у логіко-лінгвістичній моделі, що поєднує декілька простих предикатів першого порядку.

Тип 1. Логіко-лінгвістична модель складного речення природної мови, у якому за допомогою протиставних сполучників "а", "але", "проте", "однак" або часток "тільки", "лише", "же" оформлені відношення зіставлення або проставлення явищ і подій, або відношення не одночасності явищ, а їх чергування або виключення одного при можливій наявності іншого за допомогою розділових сполучників "то…то", "чи…чи", "або…або", матиме вигляд набору простих предикатів, поєднаних логічною операцією диз'юнкції:

$$L^{S_i} = L^{S_i}_{p^{(1)}}\left(x_i^{(1)}, g_i^{(1)}, y_i^{(1)}, q_i^{(1)}, z_i^{(1)}, r_i^{(1)}, h_i^{(1)}\right) \vee \ldots \vee$$
$$L^{S_i}_{p^{(\lambda)}}\left(x_i^{(\lambda)}, g_i^{(\lambda)}, y_i^{(\lambda)}, q_i^{(\lambda)}, z_i^{(\lambda)}, r_i^{(\lambda)}, h_i^{(\lambda)}\right) \vee \ldots \vee$$
$$L^{S_i}_{p^{(\varphi)}}\left(x_i^{(\varphi)}, g_i^{(\varphi)}, y_i^{(\varphi)}, q_i^{(\varphi)}, z_i^{(\varphi)}, r_i^{(\varphi)}, h_i^{(\varphi)}\right),$$

де $L^{S_i}_{p^{(\lambda)}}\left(x_i^{(\lambda)}, g_i^{(\lambda)}, y_i^{(\lambda)}, q_i^{(\lambda)}, z_i^{(\lambda)}, r_i^{(\lambda)}, h_i^{(\lambda)}\right)$ – предикатний вираз, що описує частину складного речення S_i, яка відображає закінчений зміст, виражається за допомогою простого речення рівня вкладеності $\lambda = \overline{1,\varphi}$; i – номер речення природної мови у тексті; φ – кількість простих речень у складному.

Наприклад, для речення природної мови "*Від цього вугілля розжеврювалося, займалося вогнем, червоно палаххотіло, а довкола злітали й потріскували іскри*" (*Мешак Азаре*) логіко-лінгвістична модель типу 1 матиме вигляд:

$$L^{S_1} = [p_{11}(x_1,0,0,0,0,0,0) \,\&\, p_{12}(x_1,0,0,0,0,0,0) \,\&\, p_{13}(x_1,0,0,0,0,0,h_1)] \vee$$
$$[p'_{11}(x'_1,0,0,0,0,0,h'_1) \,\&\, p'_{12}(x'_1,0,0,0,0,0,h'_1)].$$

$$L^{S_1} = [\textit{розжеврювалося}(\textit{вугілля},0,0,0,0,0,0)\&$$
$$\textit{займалося}(\textit{вугілля},0,\textit{вогнем},0,0,0,0)\&$$
$$\textit{палахкотіло}(\textit{вугілля},0,0,0,0,0,\textit{червоно})] \vee$$
$$[\textit{злітали}(\textit{іскри},0,0,0,0,0,\textit{довкола})\&\ \textit{потріскували}(\textit{іскри},0,0,0,0,0,\textit{довкола})].$$

Це складносурядне речення із відношенням зіставлення, для чого використано сполучник "а", загальна кількість простих речень у ньому – п'ять, представлення однорідних членів, зокрема присудків у кожному простому реченні інтерпретовано за допомогою логічної операції кон'юнкції, а відношення зіставлення між простими реченнями у структурі складного – через операцію диз'юнкції.

Тип 2. Формальне представлення складного речення природної мови, частини якого рівноправні за змістом, представляє собою набір простих предикатів, поєднаних логічною операцією кон'юнкції. При цьому потужність множини відношень та множини суб'єктів відношень повинна бути не менше двох, а логіко-лінгвістична модель матиме вигляд:

$$L^{S_i} = L^{S_i}_{p^{(1)}}\left(x_i^{(1)}, g_i^{(1)}, y_i^{(1)}, q_i^{(1)}, z_i^{(1)}, r_i^{(1)}, h_i^{(1)}\right)\&\ldots\&$$
$$L^{S_i}_{p^{(\lambda)}}\left(x_i^{(\lambda)}, g_i^{(\lambda)}, y_i^{(\lambda)}, q_i^{(\lambda)}, z_i^{(\lambda)}, r_i^{(\lambda)}, h_i^{(\lambda)}\right)\&\ldots\&$$
$$L^{S_i}_{p^{(\varphi)}}\left(x_i^{(\varphi)}, g_i^{(\varphi)}, y_i^{(\varphi)}, q_i^{(\varphi)}, z_i^{(\varphi)}, r_i^{(\varphi)}, h_i^{(\varphi)}\right),$$

де $L^{S_i}_{p^{(\lambda)}}\left(x_i^{(\lambda)}, g_i^{(\lambda)}, y_i^{(\lambda)}, q_i^{(\lambda)}, z_i^{(\lambda)}, r_i^{(\lambda)}, h_i^{(\lambda)}\right)$ – предикатний вираз, що описує частину складного речення S_i, яка відображає закінчений зміст, виражається за допомогою простого речення рівня вкладеності $\lambda = \overline{1,\varphi}$; i – номер речення природної мови у тексті; φ – кількість простих речень у складному.

Такий спосіб зв'язку може бути реалізований за допомогою різноманітних синтаксичних конструкцій, кожна з яких відповідно до розроблених автором правил [12] вносить відповідні зміни до простих предикатів логіко-лінгвістичної моделі складного речення.

Складні речення можна представити у вигляді окремої форми логіко-лінгвістичної моделі типу 2, якщо:

2.1. Прості речення у складному пов'язані за допомогою сурядних сполучників.

2.2. Формалізується складнопідрядне означальне речення, у якому друга частина відноситься до суб'єкта або об'єкта головного речення і виражає його ознаку.

2.3. Підрядне речення приєднане до головного сполучниками "що", "ніби", "як", "мовби" та інші, а зміст пояснювального слова (іменника або прислівника) розкривається у підрядному з'ясувальному реченні.

2.4. Декілька підрядних з'ясувальних речень приєднані до одного головного речення різними сполучниками і сполучними словами або синтаксична конструкція передає пряму мову непрямою.

2.5. У головному реченні присутні сполучні слова "там", "туди", "звідти", з якими співвідносяться сполучні слова підрядного речення місця "де", "куди", "звідки".

Так, логіко-лінгвістична модель для складного речення природної мови "*Батько витягує посудину з рідиною з вогню, несе її обережно, наче кицька своє кошеня, і старанно розливає в глиняні форми*" матиме вигляд:

$$L^{S_1} = p_{11}(x_1,0,y_1,0,z_1,0,h_{11}) \,\&\, p_{12}(x_1,0,y_1,0,0,0,h_{12}) \,\&\, p_{12}(x_1',0,y_1',q_1',0,0,h_{12}) \,\&\, p_{13}(x_1,0,z_1,0,z_1',r_1',h_{13}).$$

$$L^{S_1} = \textit{витягує(батько,0,посудину,0,рідиною,0,вогню)\&}\ \textit{несе(батько,0,посудину,0,0,0,обережно)\&}\ \textit{несе(кицька,0,кошеня,своє,0,0,обережно)\&}\ \textit{розливає(батько,0,форми,глиняні,0,0,старанно).}$$

Побудована логіко-лінгвістична модель відображає зміст складнопідрядного пояснювального речення зі сполучником "наче", де дії відбуваються одночасно. Саме тому таке речення природної мови потрібно віднести до другого типу окремої форми логіко-лінгвістичної моделі.

Тип 3. Відображає типові зв'язки в підрядному реченні природної мови, у якому йдеться про умови, мету, причину, допуск, наслідок та ін. і для якого головна та залежна частини з'єднуються в логіко-лінгвістичній моделі логічною операцією імплікації "$\rightarrow$", при цьому потужність множини відношень та множини суб'єктів відношень повинні бути не менше двох, а логіко-лінгвістична модель набуває вигляду:

$$L^{S_i} = L^{S_i}_{p^{(\lambda)}}\left(x_i^{(\lambda)}, g_i^{(\lambda)}, y_i^{(\lambda)}, q_i^{(\lambda)}, z_i^{(\lambda)}, r_i^{(\lambda)}, h_i^{(\lambda)}\right) \rightarrow L^{S_i}_{p^{(\lambda+1)}}\left(x_i^{(\lambda+1)}, g_i^{(\lambda+1)}, y_i^{(\lambda+1)}, q_i^{(\lambda+1)}, z_i^{(\lambda+1)}, r_i^{(\lambda+1)}, h_i^{(\lambda+1)}\right),$$

де $L^{S_i}_{p^{(\lambda)}}\left(x_i^{(\lambda)}, g_i^{(\lambda)}, y_i^{(\lambda)}, q_i^{(\lambda)}, z_i^{(\lambda)}, r_i^{(\lambda)}, h_i^{(\lambda)}\right)$ – предикатний вираз, що описує частину складного речення S_i, яка відображає закінчений зміст,

виражається за допомогою простого речення; i – номер речення природної мови у тексті.

Складні речення можна представити у вигляді окремої форми логіко-лінгвістичної моделі типу 3, якщо:

3.1. Підрядні речення часу, у яких присутні сполучники, що виражають відношення співвіднесеності дій підрядного і головного речень, наприклад, "після того як", "щойно", "ледве" та ін.

3.2. Підрядне речення умовне і виражає умову, за якої можлива дія, про яку йдеться у головному реченні, та приєднується сполучниками "якщо", "якби", "тоді" та ін.

3.3. Підрядне речення причини та мети, виражає причину дії, про яку говориться у головному реченні, та приєднується сполучниками "бо", "тому що", "внаслідок того що", "для того щоб" та ін. або підрядне речення наслідкове та виражає наслідок дії.

Для складного речення природної мови «*Маленький хлопчик завзято працює ковальським міхом, а батько тим часом готує форми, щоб залити в них розтоплену мідь*» логіко-лінгвістична модель матиме вигляд:

$$L^{S_1} = p_1(x_1, g_1, y_1, q_1, 0, 0, h_1) \,\&\, p_1'(x_1', 0, y_1', 0, 0, 0, h_1') \rightarrow$$
$$p_1''(0, 0, y_1', 0, z_1'', r_1'', 0).$$

$$L^{S_1} = працює(хлопчик, маленький, міхом, ковальським, 0, 0, завзято) \,\&$$
$$готує(батько, 0, форми, 0, 0, 0, тим\_часом) \rightarrow залити(0, 0, форми, 0, мідь, розтоплену, 0).$$

Дане речення складається з трьох простих та являється комбінацією сурядного зв'язку, для чого використовується сполучник "а" та підрядного причини з сполучником "щоб", тому у логіко-лінгвістичній моделі вжито дві логічні операції: кон'юнкція та імплікація.

Висновки. У матеріалах статті запропонована класифікація окремих форм логіко-лінгвістичних моделей залежно від типу речення природної мови. Оскільки під кожною компонентою логіко-лінгвістичної моделі розуміється конкретне слово, а не лише його формальне позначення, це дає змогу створити шаблон для різних синтаксичних конструкцій, зокрема і для складних сурядних та підрядних речень природної мови. Такі шаблони з певними логічними операціями, залежно від знаків пунктуації та сполучних слів, що вжиті у реченнях, а також специфічний порядок з'єднання простих предикатів, залежно від логічних зв'язків між простими реченнями у структурі складного, дають можливість розробити алгоритм для подальшого аналізу текстової інформації та її відновлення із формальних моделей представлення знань.

Список літератури:
1. *Вавіленкова А.І.* Аналіз і синтез логіко-лінгвістичних моделей речень природної мови: монографія / *А.І. Вавіленкова*. – К.: ТОВ "СІК ГРУП Україна", 2017. – 152 с.
2. *Ланде Д.В.* Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки: навч. посіб. / *Д.В. Ланде*, *І.Ю. Субач*, *Ю.Є. Бояринова*. – К.: ІСЗЗІ КПІ ім. Ігоря Сікорського, 2018. – 300 р.
3. *Широков В.А.* Язык. Информация. Система: Трансдисциплинарность в лингвистике / *В.А. Широков*. – К., 2017, – 280 с.
4. *Zhang Y.* Discriminative syntax-based word ordering for text generation/ *Y. Zhang* // Computational linguistics. – 2015. – Vol. 41. – P. 503-538.
5. *Che W.* Deep learning in lexical analysis and parsing / *W. Che, Y. Zhang* // Deep learning in Natural Language Processing, Springer Nature Singapure Pte Ltd. – 2018. – ch.4. http:// doi.org/10.1007/978-981-10-5209-5_4.
6. *Gatterbauer W.* Towards domain-independent information extraction from web tables / *W. Gatterbauer*, *P. Bohunsky*, *M. Herzog*, *B. Krupl, B. Pollak* // Proceedings WWW-07, Banff, Canada. – 2007. – P. 71–80.
7. *Briggs J.* The Ultimate Performance Metric in NLP / *J. Briggs*. – 2021, режим доступу: https://towards-datascience.com/the-ultimate-performance-metric-in-nlp-111df6c64460 (дата звернення 03.03.2021).
8. *Khairova N.F.* The Logical-Linguistic Model of Fact Extraction from English Texts / *N.F. Khairova, S. Petrasova, A.P.S. Gautam* // Communications in Computer and Information Science, Springer, Cham. – 2016. – Vol. 639. https://doi.org/10.1007/978-3-319-46254-7_51.
9. *Khairova N.* Logical-linguistic model for multilingual Open Information Extraction / *N. Khairova, O. Mamyrbayev, K. Mukhsina, A. Kolesnyk* // Cogent Engineering. – 2021. doi: 10.1080/23311916. 2020.1714829.
10. *Vavilenkova A.* Modelling of the context links between the natural language sentences / *A. Vavilenkova* // Proceedings of the 9th International Scientific and Practical Conference "Information Control Systems & Technologies" (ICST2020). – 2020. – P. 282-293.
11. *Конверський А.Є.* Сучасна логіка (класична та некласична) / *А.Є. Конверський*. 2-ге вид. перероб. та доп. – К.: Центр учбової літератури. – 2017. – 294 с.
12. *Vavilenkova A.* Basic principles of the synthesis of logical-linguistic models / *A.Vavilenkova* // Cybernetics and systems analysis. – 2015. – Vol. 51(5). – P. 826-834. http:// doi.org/10.1007/s10559-015-9776-z.

References:
1. Vavilenkova, A. (2017), *Analysis and Synthesis of logic and linguistic models for natural language sentences,* TOV "SIK GROUP UKRAINE", Kyiv, 152 p.
2. Lande, D.V., Subach, I.Y. and Boyarinova, Y.E. (2018), *Fundamentals of Theory and Practice of Intelligent Data Analysis in Cybersecurity*, ISZZI KPU, Kyiv, 300 p.
3. Shirokov, V.A. (2017), *Language. Information. System: Transdisciolinarity in linguistics*, Kyiv, 280 p.
4. Zhang, Y. (2015), "Discriminative syntax-based word ordering for text generation", *Computational linguistics*, Vol. 41, pp. 503-538.
5. Che, W. and Zhang, Y. (2018), "Deep learning in lexical analysis and parsing" in *Deep learning in Natural Language Processing*, Springer Nature Singapure Pte Ltd., ch.4. http:// doi.org/10.1007/978-981-10-5209-5_4.
6. Gatterbauer, W., Bohunsky, P., Herzog, M., Krupl, B. and Pollak, B. (2007), "Towards domain-independent information extraction from web tables", *Proceedings WWW-07*, Banff, Canada, pp. 71-80.

7. Briggs, J. (2021) "The Ultimate Performance Metric in NLP", available at: https://towardsdatascience.com/the-ultimate-performance-metric-in-nlp-111df6c64460 (accessed 3 March 2021).
8. Khairova, N.F., Petrasova, S. and Gautam A.P.S. (2016), "The Logical-Linguistic Model of Fact Extraction from English Texts", *Communications in Computer and Information Science*, vol 639. Springer, Cham. https://doi.org/10.1007/978-3-319-46254-7_51.
9. Khairova, N., Mamyrbayev, O., Mukhsina, K. and Kolesnyk, A. (2020), "Logical-linguistic model for multilingual Open Information Extraction", *Cogent Engineering*, doi: 10.1080/23311916.2020.1714829.
10. Vavilenkova, A. (2020), "Modelling of the context links between the natural language sentences", *Proceedings of the 9th International Scientific and Practical Conference "Information Control Systems & Technologies" (ICST2020)*, pp. 282-293.
11. Konverskiiy, A.E. (2017), *Modern logic (classic and non-classic)*, The Centre of Educational Literature, Kyiv, 294 p.
12. Vavilenkova, A. (2015), "Basic principles of the synthesis of logical-linguistic models", *Cybernetics and systems analysis*, Vol. 51(5), pp. 826-834, http:// doi.org/10.1007/s10559-015-9776-z.

Статтю представив д-р техн. наук, проф. Черкаського національного університету ім. Б. Хмельницького Голуб С.В.

Надійшла (*received*) *23.03.2021*

Vavilenkova Anastasiia, Dr.Sci.Tech, DSc, Docent, Professor
National Aviation University
Ave. Liubomira Guzara, 1, Kyiv, Ukraine, 03058
Tel: (066) 751-65-01, e-mail: vavilenkovaa@gmail.com
ORCID ID: 0000-0002-9630-4951

УДК 510.635:004.891(045)

Взаємозв'язок типів окремих форм логіко-лінгвістичних моделей та видів речень природної мови / Вавіленкова А.І. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2021. – № 1 (5). – С. 77 – 85.

Запропоновано апарат логіко-лінгвістичного моделювання для вилучення знань з електронних текстових документів. Виокремлено різні типи форм логіко-лінгвістичних моделей для формалізованого представлення складних речень природної мови. Сформульовано правила та здійснено класифікацію окремих форм логіко-лінгвістичних моделей залежно від синтаксичних конструкцій речень природної мови. Іл.: 1. Бібліогр.: 12 назв.

Ключові слова: логіко-лінгвістична модель; вилучення знань; електронний текстовий документ; окрема форма; речення природної мови.

УДК 510.635:004.891(045)

Взаимосвязь отдельных форм логико-лингвистических моделей с видами предложений естественного языка / Вавиленкова А.И. // Вестник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2021. – № 1 (5). – С. 77 – 85.

Предложен аппарат логико-лингвистического моделирования для извлечения знаний из электронных текстовых документов. Выделены различные типы форм логико-лингвистических моделей для формализованного представления сложных предложений естественного языка. Сформулированы правила и осуществлена классификация отдельных форм логико-лингвистических моделей в зависимости от синтаксических конструкций предложений естественного языка. Ил.: 1. Библиогр.: 12 назв.

Ключевые слова: логико-лингвистическая модель; извлечение знаний; электронный текстовый документ; отдельная форма; предложение естественного языка.

UDC 510.635:004.891(045)

Interrelation of types of separate forms of logic and linguistic models and types of sentences of natural language / Vavilenkova A.I. // Herald of the National Technical University "KhPI". Series of "Informatics and Modeling". – Kharkov: NTU "KhPI". – 2021. – № 1 (5). – P. 77 – 85.

The article proposes an apparatus of logic and linguistic modeling for extracting knowledge from electronic text documents. Different types of forms of logic and linguistic models have been singled out for the formalized representation of complex sentences of natural language. The author formulates the rules and carries out the classification of separate forms of logic and linguistic models depending on syntactic constructions of sentences natural language. Such separate forms of logic and linguistic models are templates with certain logical operations, which are used depending on the punctuation marks and connecting words applied in sentences. The specific order of simple predicates depends on the logical connections between simple sentences in the structure of the complex one. This allows to develop an algorithm for further analysis of textual information and its recovery from formal models of knowledge representation. Figs.: 1. Refs.: 12 titles.

Keywords: logic and linguistic model; knowledge extraction; electronic text document; separate form; sentence of natural language.