Evaluating the Influence of Argumentation Markers on the Identification of Reasoning Models

I.S. Pimenov[1[0000-0001-5946-9469]] and N.V. Salomatina[2[000001-0001-2345-6789]]

[1] Novosibirsk State University, 1 Pirogova, Novosibirsk, Russia
[2] Sobolev Institute of Mathematics, 4 Koptuga, Novosibirsk, Russia
pimenov.1330@yandex.ru

Abstract. The article focuses on evaluating the influence of argumentation markers on the identification of specific reasoning models with machine learning methods. The evaluation process consists of a sequence of classification experiments with different feature sets. The experiments cover the identification of arguments with three specific reasoning models: “Expert Opinion”, “Example”, and “Practical Reasoning”. These models are characterized by 1) an active use in scientific articles (as evidenced by their high frequency in the employed corpus) and 2) reliance of their textual expression on typical words and phrases (markers). Each model corresponds to a separate subset of the overall dataset: 680 arguments for classifying the “Example” model, 386 for “Practical Reasoning”, 172 for “Expert Opinion” (in each case, a half of the arguments employs the corresponding model, while the other half relies on any other model except for these three). The overall dataset contains 1975 arguments from 45 scientific articles in Russian language (on linguistics and computational technologies). The argumentation in these articles is annotated with the ArgNetBank Studio platform. Classification experiments employ machine learning methods of different types: multinomial naive Bayes, support vector machine, and multilayer perceptron. The feature sets differ by the inclusion or exclusion of discourse markers and persuasion modes indicators (expressions characterizing three argumentation aspects: logos, pathos, and ethos). The experiments show that the best improvement of identification scores (on average across all schemes and classifiers) corresponds to the representation of arguments with discourse markers (plus 10% for precision and 7% for F-measure over the lemmas baseline).

Keywords: Argument mining, reasoning models, machine learning, discourse markers, argumentation markers, scientific articles.

1 Introduction

Automatic identification of arguments in texts becomes a prerequisite step for the diverse practical applications. These uses include, for instance, the analysis of texts persuasiveness (in tasks such as the assessment of scientific articles in the aspect of their conclusions justification). The identification of arguments also enables the detection of specific types of premises for analyzed theses (why authors of a scientific article apply a certain method to the task, or choose one algorithm over others, etc.).

The use of the supervised machine learning methods (ML) requires the availability of text corpora with annotated argumentation. One of the questions during the argumentation annotation concerns the inclusion of discourse markers within the boundaries of premises and conclusions (whether such expressions as "if..., then..." or "according to..." should be annotated along with the propositional content of statements or separated from it), particularly in case of corpora intended for ML methods training. Experiments in the presented research address the necessity of such an inclusion.

The article describes experiments in identifying the semantic types of arguments at the level of reasoning models (also called argumentation schemes). This task corresponds to the separate and final stage in the pipeline of extracting argumentation structures. The aim of the work is to evaluate the influence of argumentation markers on the identification of reasoning models with machine learning methods. As such, optimization of identification scores across feature sets lies beyond the scope of the article. We calculate the lower bounds for the schemes identification (baseline) without using Deep Learning models, as the available corpus does not contain enough arguments at the moment.

2 Related works

The task of automatic identification of argumentation schemes (in the general framework of argumentation extraction) is first addressed in works [1] and [2]. The author of the first, Douglas Walton, suggests a six-stage approach to identifying arguments and reasoning models in their organization. The approach consists in first detecting arguments in a text, and then classifying them over the list of specific schemes.

The traditional approach to different stages of extracting arguments relies on employing different ML methods (DL, as of late). Vector representation of arguments for ML can characterize them in various aspects, particularly by usage of the argumentation markers. However, works with marker-based representation of arguments differ in evaluation of the markers influence on the extraction quality.

The article [3] provides an example of analyzing the influence of features of different types on the quality of extracting argumentative sentences. The authors demonstrate that the exclusive use of discourse markers (and analogical constructions for improving text coherency) without other feature types does not yield satisfying results: in the experiment, accuracy for the Maxent classifier reaches only 57.98%. In turn, combining markers and unigrams results in classification accuracy exceeding 70%. The work does not specify the classification accuracy for unigrams and bigrams in absence of markers (in case when discourse markers are excluded from annotated arguments).

The work [4] describes an experiment in context-based classification of text segments into argumentative and non-argumentative by using a BERT model and a fully connected neural network. The experiment employs a corpus of popular science articles in Russian language. The authors show that the explicit marking of argumentation markers results in a moderate increase of the classification precision (by 1%) and recall (by 5%).

The paper [5] describes a combined approach to identifying distinct elements of the argumentation structures. The authors employ discourse markers for the detection of argumentative connections (to check whether adjacent propositions in a text are related in the structure of reasoning). While the use of markers reaches the precision of 89%, the recall equals only 4% due to the low frequency of the markers in the dataset (and the need of supplementing them with other features). In turn, for identifying the exact argumentation schemes the authors employ a Naïve Bayes classifier with features of diverse types (unigrams, bigrams, part-of-speech tags, punctuation signs). The experiment covers the identification of two argumentation schemes ("Expert Opinion" and "Positive Consequences"), where the average precision across proposition types reaches respectively 87% and 80%, while the average recall equals 81% and 67%. The article does not contain a specific evaluation of the markers influence on identifying argumentation schemes.

The authors of [6] analyze the applicability of discourse markers for the automatic identification of argumentative relations in scientific papers (in biomedical domain). They employ a set of regular expressions for more than 100 discourse indicators (both separate words and compound phrases). However, the experiment shows that the use of discourse markers results in a decrease of identification quality from the baseline approach (based on the textual intersection processing). After analyzing the identification errors, the authors suggest that discourse markers in scientific articles do not necessarily organize the relevant reasoning, but instead frequently express the decorative (rhetoric) function in non-argumentative contexts.

The authors of [7] employ a multi-class SVM classifier for identifying argumentative roles of text segments in student essays, as well as for detecting arguments in support of a given thesis. They analyze the applicability of specific features from a composite set with various structural, lexical, syntactic, and contextual characteristics, as well as markers of different types (discourse and temporal markers, personal and possessive pronouns). The experiment demonstrates that the F1 value for marker-based classification of argumentative roles ranges from 26.5% to 73% (depending on the role). Addressing the task of detecting arguments in support of a given thesis, the authors arrive at conclusion that while the separate use of markers is less efficient than the separate use of lexical and syntactic characteristics, the combination of both feature types achieves the best results.

In [2], Feng and Hirst develop a similar method to [1] based on classifying arguments by their schematic structures (for the five most frequent models in their dataset). Their study focuses on automatic classification of arguments with five frequent schemes ("Example", "Cause to Effect", "Practical Reasoning", "Positive / Negative Consequences", "Verbal Classification"). The classification is approached as a separate step in the pipeline (with the assumption that arguments have been extracted on the previous step). The dataset contains 393 arguments overall (from 41 to 149 for a specific scheme). The feature set combines general features for all schemes (seven positional characteristics, such as the relative position of the conclusion and the premise, the length of the interval between them in the text) and scheme-specific features (which range from keywords and punctuation signs to the syntactic dependency relations). The authors employ the decision tree algorithm in two classification modes: one scheme against others and binary across scheme pairs. They demonstrate the significant dependence of the classification quality on the analyzed scheme: the best average accuracy reaches 90% for “Example” and “Practical Reasoning”, but only 70% for “Cause to Effect” and 60% for “Positive / Negative Consequences” and “Verbal Classification” (least represented in the dataset, with 44 and 41 arguments). Features that are specific to particular schemes effectively correspond to markers of these schemes, but the article does not address the influence of these features.

The authors of [8] address the identification of “Expert Opinion” arguments in texts in Russian language. The identification employs lexical-grammatical patterns that are constructed by experts. These patterns correspond to specific combinations of discourse connectives, verbs and nouns of diverse semantic classes, as well as their integrating constructions (templates with variables). Constants in the templates correspond to markers. The precision of the identification reaches 86.5%.

Overall, the existing works in automatic identification of arguments with specific reasoning schemes focus prevalently on texts in English language. One existing work that addresses the task for texts in Russian language ([8]) limits the scope to just one type of reasoning schemes ("Expert Opinion"), and the identification of its arguments relies on expert patterns (which require extensive labor for construction and do not support the identification of arguments with other schemes). Another known work [9] addresses the automatic extraction of arguments at the level of their stance (supporting or attacking a given thesis).

3 Identification of Reasoning Models as a Classification Task

We employ a pipeline-based approach to extracting argumentation structures from texts and, in the present article, focus on the final stage: the identification of specific reasoning models. The input at this stage corresponds to arguments (sets of detected argumentative statements with established connections between them). We have addressed the preceding stages in our earlier works ([10], [11]).

Let $C = \{t_i\}$ denote a corpus of texts, $0 < i \leq I$, where I is the number of texts in the corpus. $A^i = \{a_j^i\}$ is the set of arguments in the text t_i ($0 < j \leq J$, J is the number of arguments in this text). An argument a_j^i consists of its forming statements (u_j^i) and their connecting reasoning scheme (sch_j^i): $a_j^i = \{u_j^i, sch_j^i\}$ ($u_j^i = \{\{p_{jk}^i\}, c_j^i\}$, where $\{p_{jk}^i\}$ is the set of premises ($k > 0$) in the argument a_j^i, while c_j^i is its conclusion). In this article we focus on three specific argumentation schemes: Sch = {Expert Opinion, Example, Practical Reasoning}. These schemes are characterized by 1) an active use in scientific articles of the chosen thematic areas (information technologies and linguistics), and 2) their frequent expression in texts with explicit markers of diverse types.

Fig. 1 provides an example of an argumentation graph fragment with three arguments implementing the analyzed schemes. Arguments A33 and A29 support the same conclusion (S36) with different premises (S37 and S20) connected to this conclusion by different argumentation schemes. The statement S20 serves as a premise in A29 and as a conclusion in A17. The text of statements has been translated from Russian.

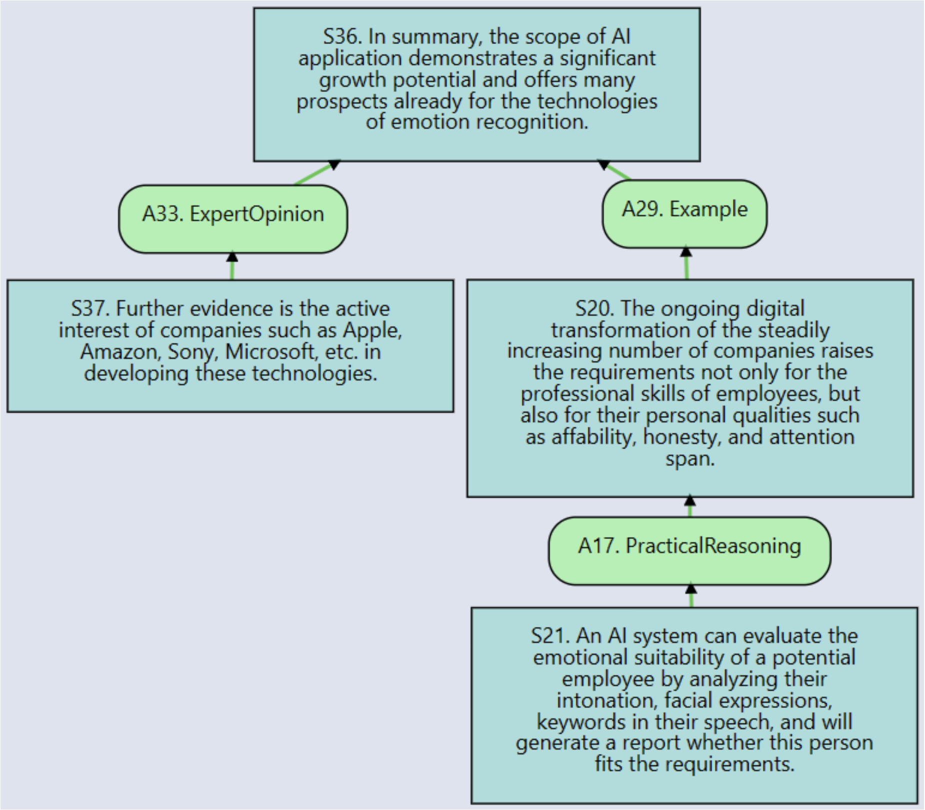
S36. In summary, the scope of AI application demonstrates a significant growth potential and offers many prospects already for the technologies of emotion recognition.
A33. ExpertOpinion
A29. Example
S37. Further evidence is the active interest of companies such as Apple, Amazon, Sony, Microsoft, etc. in developing these technologies.
S20. The ongoing digital transformation of the steadily increasing number of companies raises the requirements not only for the professional skills of employees, but also for their personal qualities such as affability, honesty, and attention span.
A17. PracticalReasoning
S21. An AI system can evaluate the emotional suitability of a potential employee by analyzing their intonation, facial expressions, keywords in their speech, and will generate a report whether this person fits the requirements.

Fig. 1. An example of an argumentation graph fragment with the three analyzed schemes.

We address the task of binary classification for each a_j^{ti}: whether the analyzed argument corresponds to the reasoning through the given argumentation scheme.

4 Methods for Identification of Reasoning Models

4.1 Classification by Machine Learning Algorithms

The base representation of an argument $\{\{p_{jk}^i\}, c_j^i\}$ (its textual expression) corresponds to a vector of lemmas containing at least one Cyrillic symbol. The selection of lemmas for the base feature set relies on the χ^2 criterion: this criterion enables the filtering of lemmas whose distribution across arguments does not provide an informative indication of argumentation schemes. We assign the χ^2 threshold in accordance with the empirical observations.

The experiments employ three classification algorithms belonging to different functional types and frequently used in Argument Mining research. These algorithms are

SVM (support vector machine), MNB (multinomial Naïve Bayes), and MLP (multilayer perceptron, a basic neural network). We use program realizations of the algorithms from the Scikit-Learn library for machine learning in Python [12].

The SVM algorithm. The training of the algorithm relies on the RBF kernel (Radial Basis Function). Two key parameters for this kernel are C (which regulates the balance between classification errors and simplicity of the training surface) and γ (which defines the influence of a single training example). We assign their values empirically: C = 100, γ = 1 / (n_features * var), where n_features is the number of features, and var is the variance of the training matrix. The choice of the RBF over other kernel functions is based on preliminary experimental results: as the classification reached similar scores for RBF and other kernel types, we have chosen the former due to it being the default kernel function in the employed implementation of the algorithm.

The MLP algorithm. The training of the neural network consists in assignation of weights for its constituting neurons by the backpropagation method. The network contains 100 neurons in the hidden layer and uses the logistic sigmoid function for its activation. Iterations continue until convergence (when the score or loss are not increasing by more than 1e-4 for 10 iterations) or until reaching the maximum number of iterations (200). Additionally, an early stop check analyzes the validation data (10% of the training examples) to avoid overfitting. The regularization parameter α equals 0.0001.

The MNB algorithm. To choose a class label for an input vector, the classifier evaluates the probability of this vector belonging to each of the possible classes with the Bayes' theorem. The Laplace smoothing parameter (α = 1) balances the incompleteness of the training set by preventing the assignation of zero probability to the absent features.

4.2 Two Types of Argumentation Markers

The study investigates the influence of two types of argumentation markers on reasoning models identification. The first type denotes discourse markers that organize a text as the coherent unit (including the level of transitions between argument components, from premises to conclusions). The second type corresponds to indicators of three persuasion modes (aspects of the argumentative effect): persuasion by logical facts (logos), by emotional manipulation (pathos), or by appeals to a source of authority (ethos).

Discourse markers. We extract markers from a corpus of texts with expert-annotated argumentation, and then expand their list with separate rhetoric markers from the RSTreeBank resource (https://rstreebank.ru/), as well as with marker synonyms from various online synonym dictionaries. The resulting dictionary of discourse markers contains 407 words and word combinations. Some of the markers include punctuation signs, begin strictly with the capitalized letter (to indicate their position at the start of a sentence), or correspond to a shortened form. These specifications serve to improve the model identifications. Examples of such markers are "Tak," ("So,"), "Poetomu" ("That is why"), "napr." (a shortened version of "naprimer", which means "for example", corresponds to "e.g.").

Indicators of ethos, pathos, and logos. According to the classic model by Aristotle, argumentation achieves a persuasive effect through three different aspects: intellectual (logos), emotional (pathos), and authoritative (ethos). These three aspects correspond to three different modes of influencing the audience. Textual expression of the modes relies on specific constructions with lexical units that correspond to each mode.

Identification of persuasion modes in argumentative statements presupposes the creation of dictionaries with indicators that mark these modes. In turn, specification of indicator forms needs to follow explicitly formulated criteria, such as given in [11].

Indicators of *pathos* correspond to at least one of the three following properties:

1) *Redundancy*: a language expression does not contribute to presentation of information (neither to the semantic content nor the structure), and its removal does not influence the statement neither in content nor in relations to other statements. Examples of such indicators are "*lish*" ("only", "merely"), "*dazhe*" ("even"), "*vovse*" (an emphatic particle with different contextual meanings, such as "at all" or "completely").

2) *Deontic modality*: an expression conveys a meaning of prescribing an action ("*trebuetsa*" ("it is required to"), *nuzhno* ("it is necessary to"), "*vazhno*" ("it is important to"), etc.).

3) Stylistic marking: an expression can be substituted by a stylistically neutral synonym without affecting the meaning of the statement ("*vpechatlyaushij*" ("impressive"); "*vydajushijsya"* ("remarkable")).

In the presented research, we employ a broad definition of ethos: justification of a claim through backing it with a source of authority. This source can correspond to a specific individual (a named scientist, an expert), or a group (such as a research team), or an impersonal agent (a popular opinion, uncertain informer), or an applicable example case (a precedent, a traditional practice). Consequently, indicators of *ethos* possess the property of *authorization*: they indicate a source of information ("*experty chitajut*" ("experts think that"), "*po mneniju avtora*" ("in the author's opinion")). This definition also includes bibliographic references, usually denoted by square or round brackets.

Finally, indicators of logos are expressions that organize the reasoning structure in its textual presentation. They contribute to *logical connectivity* of the text. Examples of logos indicators are "*esli…, to…*" (if …, then…), "*vo-pervyh*", "*vo-vtoryh*" ("first of all", "secondly").

To perform the automatic identification of indicators, we construct search patterns in form of regular expressions. An indicator may contain several elements (framed in square brackets in the pattern description) with specified lexical and/or grammatical properties of each. If an element of an indicator corresponds to several possible (alternative) constants, a delimitator "|" separates them in a list. Certain indicators permit insertions of arbitrary length (limited by the statement length), and the symbol "…" denotes these insertions. Below we present the examples of patterns for indicators of ethos (E_i), pathos (P_i), and logos (L_i).

E_i : [soglasno // PREP] [– // ADJF...(datv|loct)] [– // NOUN...(datv|loct)];
P_i : [(sovershenno|yavno|ochen') // –] [– // (ADJF|ADJS|ADVB)];
L_i : [v // PREP] [(zaklyuchenie|itog|chastnost') // -]

The first example corresponds to an expression “according to…”, where a source of authority is specified by a pair of an adjective and noun in a dative or locative case (“according to the new data”, “according to the recent results”, etc.). The second pattern specifies a combination of an emphatic particle (“completely”, “obviously”, “very”) and an adjective or an adverb. The third example denotes a connective for accentuating the beginning of a new information block (in conclusion, in particular).

At present, the indicators dictionary contains 89 patterns: 32 for ethos, 39 for pathos, 18 for logos. These numbers correspond to separate patterns (sequences of several elements) without accounting for alternative constants within indicators.

The use of indicator dictionaries in identification of argumentation schemes enables the processing of persuasion modes characteristics that are implicitly expressed in these schemes. The article [11] describes a qualitative evaluation of persuasion modes weights in argumentation schemes by their comparative functional analysis. That study groups schemes by their functional similarity and then contrasts the similar schemes within each group. At the first level, schemes are separated into two groups based on scope of the expressed reasoning. Namely, argumentation can advance either by analyzing facts within the propositional content of presented statements or by appealing to external sources of authority (which are not directly commeasurable with the analyzed phenomena). Correspondingly, the first group contains argumentation schemes with dominant *logos*, while models in the second group rely on *ethos*. However, specific schemes might complement the main persuasion component with others at different intensity.

In particular, among the schemes with the prevalent logos, abstract causal models (such as “Cause to Effect” or “Correlation to Cause”) rely more on the logical component than do practically-oriented schemes (such as “Practical Reasoning” or “Positive / Negative Consequences”). The latter models potentially convey a stronger complementary component of pathos (especially if an analysis of possible results accentuates their sentiment-based evaluation). Similarly, arguments from authority convey ethos most clearly by specifying an exact specialist (through the “Expert Opinion” model). The authoritativeness of the cited source decreases if an appeal addresses an impersonal agent (by the “Popular Opinion” scheme).

5 Classification Experiments with Different Features

5.1 Data Set with Argumentation Annotation

The experiment dataset contains 1975 arguments and 1809 argumentative statements extracted from 45 short scientific articles in Russian language with annotated argumentation. The articles range in length from 800 to 1500 words and belong to two research areas: information technologies (23 articles) and linguistics (22). Their texts have been downloaded from the freely accessible sources: online scientific library “CyberLeninka” and proceedings of the “Corpus Linguistics” conference. The expert annotation of argumentation uses tools of the ArgNetBank Studio web platform [13]. The annotated texts are available at the platform website [14]. The annotating process follows the Argument Interchange Format standard, an example of employing which for modelling argumentation in texts in Russian language is given in [15].

Two expert annotators (qualified both in linguistics and information technologies) perform the annotation of argumentation in texts. They follow a detailed annotation instruction formulated in advance. The annotation of a text consists in constructing an argumentation graph that is oriented, connected, acyclic, and rooted. The root node in each graph denotes the main thesis of the respective text. For each argument identified in a text, the annotators specify its constituents (premises and conclusions) and the semantic type of the argumentative connection between them (by indicating its argumentation scheme from Walton's compendium [16]).

For a quarter of the corpus texts (12 articles), both experts have constructed separate annotation versions. Double annotation of these texts enables the calculation of correspondence coefficients across all three levels of the argumentation structure (to ascertain the reliability of annotations). The average values across 12 texts are given below.

1) The average ratio of the number of argumentative statements, identified by both experts for the same text, to the sum of argumentative statements, identified in it by at least one, equals 78%.

2) The average ratio of corresponding connections between argument components (to the similar sum of all connections identified by both annotators) reaches 55%. This value serves as the lower bound: the same connections between same statements yet with different configurations (parallel or sequential) are considered non-corresponding.

3) The average percentage of matching argumentation schemes in connections equals 60%.

The resulting dataset contains only one argument-annotated version for each text. For texts with two annotation versions, the choice of a version follows the joint decision of the annotators. The corpus includes 330 arguments with the "Example" scheme, 193 with "Practical Reasoning", 86 with "Expert Opinion".

5.2 Construction of Training and Test Sets

For each of the three analyzed schemes, around 80% of its arguments constitute its training set (LS) for the classification, while the other 20% form the test set (TS). The exact percentage varies due to the principle of *text integrity* in dividing arguments between sets: arguments from the same text can belong either only to the LS or only to the TS (if an argument from a text belongs to the LS, all other arguments from this text can appear only in the LS, but not the TS, and vice versa). Additionally, the TS for its scheme contains an equal number of arguments from texts of both thematic fields (IT and linguistics). Negative classification examples in sets (arguments with other schemes) are extracted from the same texts as positive. They are selected at random, so that the number of negative examples equals the number of positive both in the TS and LS.

The resulting classification sets for each analyzed scheme contain the following numbers of arguments:

a) *Example*: 520 arguments in the LS (260 with "Example", 260 with other models) and 140 arguments in the TS (70 with "Example", 70 with other schemes).

b) *Practical Reasoning*: 306 arguments in the LS (153 arguments of both types) and 80 arguments in the TS (40 with "Practical Reasoning", 40 with other schemes).

c) *Expert Opinion*: 136 arguments in the LS (68 for “Expert Opinion”, 68 others) and 36 arguments in the TS (18 with “Expert Opinion”, 18 with other schemes).

The selection of lemmas for the vector representations of arguments relies on the χ^2 criterion (described in 4.1). The threshold values equal 10% for “Example” and “Practical Reasoning”, and 20% for “Expert Opinion” (established empirically). The lemmatization of words in argument components (premises and conclusions) employs the PyMorphy2 library for Python.

5.3 Comparing Classification Results across Feature Sets

We perform 10 experiments in binary classification for identifying each of the three chosen argumentation schemes. The experiments employ different feature sets. The construction of feature sets consists in different combinations of feature types in order to evaluate the influence of discourse markers and persuasion modes indicators on the identification of schemes. The number of features in each experiment depends on the employed feature type (407 discourse markers, 89 patterns of persuasion indicators) and the analyzed scheme (due to different χ^2 thresholds for the selection of lemmas: 444 lemmas for the “Example” scheme, 181 for “Expert Opinion”, 325 for ‘Practical Reasoning”).

Before the experiments on different feature sets, we perform the preliminary tests to empirically assign the threshold values for the formal filtration of features (by the χ^2 criterion). The chosen threshold values improve the average F-measure across all classification algorithms and schemes by 6.3% (over the unfiltered lemmas). The following experiments (1-11) employ the formal filtration in all cases when feature sets include lemmas.

There are 4 different types of features: lemmas without positional specification (1), discourse markers (2), persuasion indicators (3), lemmas with positional specification (4), where types (1) and (4) are mutually exclusive. The number of possible feature types combinations can be calculated with the formula $\mathrm{M} = 2 \times \sum_{k=1}^{n-1} C(n-1, k) - 3 = 14 - 3 = 11$, where $C(n-1, k)$ is the number of combinations of n – 1 elements taken k at a time. The deduction of the constant 3 is based on the mutual exclusivity of the feature types (1) and (4).

Experiment 1. Vector representations of arguments consist only of lemmas and do not include discourse markers nor persuasion modes indicators (described in 4.2 and 4.3). In effect, the initial experiment addresses the influence of thematic content of arguments on identification of each scheme.

Experiment 2. Vector representations of arguments contain only discourse markers (with the exclusion of ordinary lemmas). The classification results demonstrate the exclusive role of markers in indicating specific schemes.

Experiment 3. Vector representations of arguments contain only indicators of three persuasion modes (ethos, pathos, logos). This experiment addresses the distinguishing potential of persuasion modes indicators, which emphasize the form of arguments expression (by accentuating logical connections between facts, by invoking an emotional reaction in a reader, or by underlining authoritativeness of cited sources).

Experiment 4. Vector representation of arguments combines the features used in experiments 1 and 2. The new experiment aims at comparing the results from combining the features of different types with the quality scores for lemmas and markers when used separately (whether these scores improve, and if yes, to which degree).

Experiment 5. This experiment resembles the preceding one, but persuasion modes indicators replace discourse markers in vector representation of arguments. The comparison with the previous results will enable the evaluation of indicators efficiency in identifying schemes.

Experiment 6. Vector representations of arguments contain only persuasion modes indicators and discourse markers. The classification results will demonstrate how these two feature types strengthen or weaken each other in joint use and absence of lemmas.

Experiment 7. Vector representation of arguments combines all types of features separately employed in experiments 1, 2, and 3. The classification results demonstrate the efficiency of a multi-aspect argument modelling.

The next four experiments (8, 9, 10, 11) address the influence of the argumentative role-based distinguishing between lemmas (whether they appear in premises of an argument or the conclusion) on the identification of schemes. This is achieved by doubling the number of lemmas in the feature set (lemmas occurrence in premises and in conclusions are examined separately). The identification of roles of statements in arguments can employ machine learning methods in a similar classification task.

Experiments 8, 9, 10, 11. Each lemma in vector representation of arguments (as in experiments 1, 4, 5, 7) corresponds to two features: one specifies its occurrences in premises, another in conclusions.

The eleven experiments cover all possible combinations of the analyzed feature types. Table 1 provides the results of experiments 1–11, performed with three ML methods (described in 4.1), separately for each analyzed scheme. We employ the standard precision (P), recall (R) scores and F-measure (F) to evaluate the classification results. Experiments in table are denoted by letter “E” and their number, while supplementary labels specify the feature set composition (“L” for filtered lemmas, “M” for discourse markers, “P” for persuasion modes indicators, “R” for role specification of lemmas).

Comparing the classification results. The table shows a notable influence of a choice of a classifying algorithm on identification of schemes. For the “Example” model, the MNB algorithm achieves the best precision value across most of the experiments, while the best recall characterizes the SVM method. In turn, for “Expert Opinion” and “Practical Reasoning”, MLP demonstrates the best precision and recall both. For all three schemes across feature sets, the basic neural network (MLP) surpasses two other algorithms in identification F-measure by 8.7% on average.

Table 1. Classification quality scores for Experiments 1-11.

	Example								
	MNB			SVM			MLP		
	P	R	F	P	R	F	P	R	F
E1 (L)	0.64	0.26	0.37	0.52	0.43	0.47	0.55	0.49	0.52
E2 (M)	0.60	0.41	0.49	0.42	0.49	0.45	0.46	0.61	0.53
E3 (P)	0.67	0.29	0.40	0.71	0.21	0.33	0.61	0.24	0.35
E4 (LM)	0.65	0.29	0.40	0.57	0.49	0.52	0.53	0.43	0.47
E5 (LP)	0.74	0.24	0.37	0.60	0.30	0.40	0.52	0.39	0.44
E6(MP)	0.63	0.34	0.44	0.42	0.34	0.38	0.49	0.47	0.48
E7 (LMP)	0.63	0.31	0.42	0.42	0.34	0.38	0.50	0.53	0.51
E8 (LR)	0.82	0.26	0.39	0.54	0.44	0.49	0.58	0.44	0.50
E9 (LMR)	0.84	0.23	0.36	0.57	0.41	0.48	0.57	0.39	0.46
E10 (LPR)	0.78	0.30	0.43	0.55	0.41	0.47	0.65	0.51	0.58
E11 (All)	0.85	0.31	0.46	0.52	0.39	0.44	0.60	0.40	0.48
	Expert Opinion								
	MNB			SVM			MLP		
	P	R	F	P	R	F	P	R	F
E1 (L)	0.33	0.22	0.27	0.40	0.33	0.36	0.46	0.61	0.52
E2 (M)	0.53	0.50	0.51	0.65	0.72	0.68	0.50	0.50	0.50
E3 (P)	0.43	0.17	0.24	0.38	0.17	0.23	0.56	0.78	0.65
E4 (LM)	0.33	0.22	0.27	0.36	0.33	0.28	0.42	0.28	0.33
E5 (LP)	0.50	0.11	0.18	1.00	0.06	0.11	0.82	0.50	0.62
E6(MP)	0.64	0.50	0.56	0.64	0.50	0.56	0.64	0.39	0.48
E7 (LMP)	0.64	0.50	0.56	0.64	0.50	0.56	0.69	0.50	0.58
E8 (LR)	0.80	0.22	0.35	0.67	0.11	0.19	0.60	0.33	0.43
E9 (LMR)	0.75	0.17	0.27	0.67	0.11	0.19	0.50	0.22	0.31
E10 (LPR)	0.80	0.22	0.35	0.67	0.11	0.19	0.50	0.17	0.25
E11 (All)	0.80	0.22	0.35	0.67	0.11	0.19	0.50	0.17	0.25
	Practical Reasoning								
	MNB			SVM			MLP		
	P	R	F	P	R	F	P	R	F
E1 (L)	0.64	0.35	0.45	0.79	0.28	0.41	0.73	0.40	0.52
E2 (M)	0.52	0.28	0.36	0.45	0.50	0.48	0.50	0.53	0.51
E3 (P)	0.70	0.40	0.51	0.63	0.30	0.41	0.57	0.30	0.39
E4 (LM)	0.73	0.20	0.31	0.77	0.25	0.38	0.62	0.33	0.43
E5 (LP)	0.45	0.23	0.30	0.57	0.30	0.39	0.53	0.23	0.30
E6(MP)	0.40	0.30	0.34	0.37	0.42	0.40	0.41	0.33	0.36
E7 (LMP)	0.40	0.30	0.34	0.37	0.42	0.40	0.44	0.38	0.41
E8 (LR)	0.52	0.33	0.40	0.77	0.25	0.38	0.67	0.45	0.54
E9 (LMR)	0.50	0.30	0.37	0.75	0.23	0.35	0.54	0.38	0.44
E10 (LPR)	0.48	0.33	0.39	0.69	0.23	0.34	0.57	0.33	0.41
E11 (All)	0.50	0.30	0.37	0.60	0.15	0.24	0.48	0.35	0.41

The experiments demonstrate the significant role of markers and indicators in characterizing argument components. In E2 and E3, feature sets contain only markers or indicators respectively, yet the separate classification scores (for “Expert Opinion”, all scores) in these experiments are higher than in E1 (in representation of arguments by thematic lemmas) for 2 algorithms out of 3. The lesser efficiency of lemmas appears to be caused by thematic diversity of specialized articles in the corpus. Most of the represented themes correspond only to 2-3 articles: for example, 23 articles in IT cover a variety of subfields, such as image analysis, information security, artificial intelligence, machine translation, speech recognition, medical applications of IT, etc. As a result, thematic vocabularies differ significantly between articles. If all texts for a specific theme appear exclusively in either the LS or TS, identification scores decrease for vector representation of arguments by formally filtered lemmas. As a rule, persuasion modes indicators exhibit better precision, while discourse markers in most cases reach a better recall (for instance, as in E2 and E3 for the “Example” model).

The noted tendencies principally characterize the identification of the “Example” and “Expert Opinion” models. They apply to “Practical Reasoning” to a lesser extent, as the present incompleteness of its marker dictionary constrains its identification (the dictionary of markers for “Practical Reasoning” is at the stage of development). This difference in typical markers is caused by a specific functional trait of the “Practical Reasoning” scheme (a greater positional distance between connected statements in an argument). Namely, discourse markers specify connections mostly between adjacent statements in a text (due to their role as rhetoric connectives). The “Example” and “Expert Opinion” models operate with such adjacent statements (an example or an appeal to a source authority tends to directly succeed or precede the corresponding claim). However, in arguments based on “Practical Reasoning”, the premises and conclusions can be presented in different sections of a text (for example, one paragraph specifies the aim of a research or a separate subtask, while another describes a choice of a method in accordance with the earlier formulation of the analyzed problem). In these cases, discourse markers might not suffice for identification of “Practical Reasoning” arguments, and the use of thematic lemmas for the vector representation becomes preferable.

The analysis of classification errors across the experiments reveals three prominent causes of incorrect scheme identification. One type of errors occurs due to innate overlapping of arguments in argumentation structures, where one statement can be used a support for proving another and then in turn be justified by a third statement (within another argument). As such, the intermediary statement might contain markers relevant to the scheme of either argument (but not the other), and these markers potentially mislead the classifier. Errors of the second type stem from the appearance of ambiguous markers that either can introduce several possible schemes or correspond in general to a specific scheme, yet in a particular context are used for implementing another. The third type of errors occurs for arguments without explicit markers (either of discourse or of a persuasion mode). In these cases, classification by the exclusive use of lemmas without relying on markers can exceed in quality scores the markers-focused classification (particularly for the “Practical Reasoning” scheme).

Not always justified is the initial assumption that the combined representation of arguments by lemmas and markers (or indicators) will result in a considerable improvement of the identification scores. The E4 for "Expert Opinion" provides an example of the scores, on the contrary, decreasing. The reasons for such decreases correspond to the two outlined causes: thematic diversity of specialized articles, insufficient size of the discourse markers dictionary. The MLP classifier (best across identification scores on average) demonstrates only a 4% increase in precision when extending lemmas with indicators. Its best recall (on average across schemes) corresponds to the exclusive use of markers (4.7% higher than with lemmas). If we employ the lemmas-based classification as a baseline for comparison (on average across all algorithms and schemes), the use of discourse markers yields a 10% increase in recall, while indicators provide a 2% increase in precision. The combination of lemmas, markers, and indicators improves the precision by 6%, yet the combination only of lemmas and indicators further increases the gain to 7.3%. In terms of F-measure, markers give a 7% increase over the lemma-based baseline, while the combination of all feature types provides a boost of 3%. Finally, the elaboration of lemmas position (in a premise or a conclusion) improves the classification precision in most of cases.

6 Conclusion

The presented research focuses on a sequence of experiments to evaluate the influence of argumentation markers of different types (discourse markers and persuasion modes indicators) on identification of specific argumentation schemes with machine learning methods. The analyzed schemes correspond to arguments from “Example”, “Expert Opinion”, and “Practical Reasoning”. The binary classification employs algorithms of different functional types: SVM, MNB, and MLP. Experiments differ in feature sets for vector representation of arguments through inclusion or exclusion of various types of features (lemmas after formal filtration, discourse markers or persuasion modes indicators from the constructed dictionaries).

The experiments show that, when using markers from a dictionary of a sufficient size, the introduction of thematic lemmas into the feature set does not effectively influence the identification scores. Another observation concerns the importance of thematic homogeneity of the analyzed data (not only at the level of a general theme, but also in its sub-topics). In classifying arguments from articles in different thematic sub-fields of information technologies and linguistics, the introduction of lemmas into vectors might result in a decrease of identification scores on combined feature sets (with features of different types). As a rule, discourse markers improve the precision of identifying argumentation schemes, while persuasion modes indicators increase its recall. Precision can be further increased by specifying the positional properties of thematic lemmas (whether they occur in premises of an argument or its conclusion), yet such a specification requires the preliminary identification of statement roles in arguments.

Acknowlegment

The work was funded by Russian Science Foundation according to the research project no. 23-11-00261.

References

1. Walton, D.: Argument mining by applying argumentation schemes. Studies in Logic 4(1), 38–64 (2011).
2. Feng, V−W., Hirst, G.: Classifying arguments by scheme. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies 2011, vol. 1, pp. 987–996. Association for Computational Linguistics (ACL) (2011).
3. Moens M.-F., Boiy E., Palau R.M., Reed C. Automatic Detection of Arguments in Legal Texts. In: ICAIL '07: Proceedings of the 11th international conference on Artificial intelligence and law, June 2007, pp. 225–230.
4. Sidorova E., Akhmadeeva I. Kononenko I., Chagina P. Izvlechenie argumentatsii na osnove indikatornogo podhoda [An Indicator-Based Approach to Argumentation Mining] [In Russian]. In: Proceedings of the 20th National Conference on Artificial Intelligence KII-2022, 2022, pp. 219–233.
5. Lawrence, J., Reed, C.: Combining argument mining techniques. In: Proceedings of the 2nd Workshop on Argumentation Mining, pp. 127–136. Denver, CO, June. Association for Computational Linguistics (2015).
6. Gao, Y., Gu, N., Lam, J., Hahnloser, R.: Do Discourse Indicators Reflect the Main Arguments in Scientific Papers? In: Proceedings of the 9th Workshop on Argumentation Mining, pp. 34–50. Association for Computational Linguistics (ACL) (2022).
7. Stab C., Gurevych I. Identifying argumentative discourse structures in persuasive essays. In: Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (Doha) pp. 46–56.
8. Achmadeeva, I., Kononenko I., Salomatina N., Sidorova E.: Indicator Patterns as Features for Argument Mining. In: International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) Proceedings, pp. 0886−0891. Novosibirsk, Russia (2019).
9. Fishcheva I., Kotelnikov E. Cross-Lingual Argumentation Mining for Russian Texts. In: Analysis of Images, Social Networks and Texts, 8th International Conference, AIST 2019, Revised Selected Papers, 2019, pp. 134-144.
10. Salomatina N, Sidorova E., Pimenov I. Identification of argumentative sentences in Russian scientific and popular science texts. In: Journal of Physics: Conference Series, Volume 2099, International Conference «Marchuk Scientific Readings 2021» (MSR-2021). 2021, Russia, Novosibirsk, pp. 1–8.
11. Pimenov I., Salomatina N., Timofeeva M. The Quantitative Evaluation of the Pathos to Ethos Ratio in Scientific Texts. In: Proceedings of the 2022 IEEE 23rd International Conference of Young Professionals in Electron Devices and Materials (EDM). Altai, Russia, 2022, pp. 312–317.
12. Scikit Learn Homepage: https://scikit-learn.org/stable/supervised_learning.html, last accessed 2023/08/21.
13. Sidorova, E., Ahmadeeva I., Zagorul'ko YU., Seryj A., Shestakov V.: Research platform for the study of argumentation in popular science discourse. Ontology of Designing 10, N 4(38), 489–502 (2020).
14. ArgNetBank Studio Homepage: https://uniserv.iis.nsk.su/arg/, last accessed 2023/08/21.
15. Pimenov I.: Specifika argumentacionnogo annotirovaniya nauchnih I nauchno-populyarnih tekstov. In: Proceedings of the Corpora 2021 International Conference, pp. 330–337, Saint-Petersburg, Skifiya-Print (2021).
16. Walton D., Reed C., Macagno F.: Argumentation schemes. Cambridge University Press, New York (2008).