This paper considers an approach to resolving referential relations when extracting information from a text. The proposed approach is an attempt to integrate the multifactorial model of the activation coefficient with the approach to resolving the referential ambiguity of the text when replenishing the ontology. The found objects are compared based on an assessment of the proximity of attributes and relationships of objects. An ontological interpretation of relations and measures of similarity of attributes based on a multifactorial model is proposed. This model is distinguished by the fact that it makes it possible to introduce the concepts of "rhetorical distance", "linear distance", "animation", "distance between paragraphs", and "syntactic and semantic role of the antecedent". A multifactorial model is proposed, which is a necessary and sufficient component for the purpose of explaining the measure of similarity of referents for choosing the best applicant. The counting system and its modification were revealed by trial and error; the work was carried out until the selected numerical weights began to explain all the available material. The current study also examines the factors of choice of reference devices that make it possible to work with complex sentences and texts. Moreover, examples of finding a measure of proximity in a multilingual system for the Kazakh, Russian, and English languages are offered. For the current paper, texts in the Russian, English, and Kazakh languages were used as a source for practical tasks. The texts were selected using news articles on the Internet sites where translations into other languages, including those named above, were offered.

The authors of this study have done massive practical work, which confirms the correctness of the thesis they are considering

Keywords: information extraction, proximity measure, referential factors, semantic text analysis, anaphora

UDC 004.9

DOI: 10.15587/1729-4061.2022.255786

BUILDING A MODEL FOR RESOLVING REFERENTIAL RELATIONS IN A MULTILINGUAL SYSTEM

Yerzhan Zhumabay
Doctoral Student
Department of IT-Management
Astana International University
Kabanbay Batyra ave., 8, Nur-Sultan, Republic of Kazakhstan, 010000
Gulzhamal Kalman
Doctoral Student*
Madina Sambetbayeva
Corresponding author
PhD, Associate Professor*
Leading Researcher**
E-mail: madina_jgtu@mail.ru
Aigerim Yerimbetova
PhD, Associate Professor, Leading Researcher**
Professor***
Assem Ayapbergenova
Master of Engineering***
Almagul Bizhanova
Senior Lecturer
Department of Information Systems and Cybersecurity
Institute of Information Technologies
Almaty University of Power Engineering and
Telecommunications named after Gumarbek Daukeyev
Baitursynov str., 126/1, Almaty, Republic of Kazakhstan, 050013
*Department of Information System
L. N. Gumilyov Eurasian National University
Satpayev str., 2, Nur-Sultan, Republic of Kazakhstan, 010008
**Institute of Information and Computational Technologies
Committee of Science of the Ministry of Education and
Science of the Republic of Kazakhstan
Shevchenko str., 28, Almaty, Republic of Kazakhstan, 050010
***Department of Software Engineering
Institute of Automation and Information Technologies
Satbayev University
Satbayev str., 22 a, Almaty, Republic of Kazakhstan, 050013

Received date 02.02.2022
Accepted date 06.04.2022
Published date 30.04.2022

***How to Cite**: Zhumabay, Y., Kalman, G., Sambetbayeva, M., Yerimbetova, A., Ayapbergenova, A., Bizhanova, A. (2022). Building a model for resolving referential relations in a multilingual system. Eastern-European Journal of Enterprise Technologies, 2 (2 (116)), 27–35. doi: https://doi.org/10.15587/1729-4061.2022.255786*

1. Introduction

The integrity and consistency of the speech produced are directly related to the repeated mention of the same essences. Therefore, in speech, we often encounter the mention of certain objects of extralinguistic reality, which are called referents, and the process of referencing actualized names is referral [1]. The establishment of referential relations in discourse is one of the most pressing, but difficult to model problems of automatic text analysis. This is partly due to the possibility

of changing the meaning of the statement and the degree of its accessibility depending on which referential medium the speaker uses. It can be a proper name, a pronoun, or a descriptive name group. If one referential expression refers to another that was used earlier, then an anaphoric relation (anaphora) is established between them, the last of the expressions is called an anaphora, and the preceding one is called an antecedent.

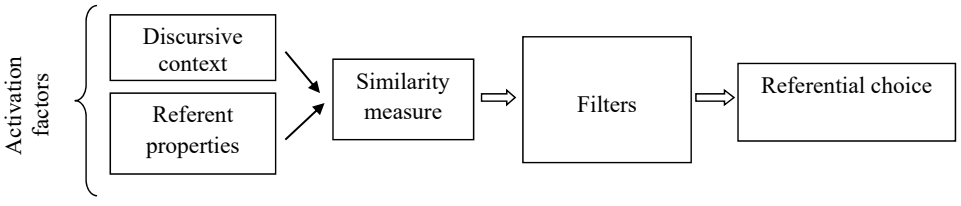

Fig. 1. Reference selection model

The reference has been the subject of a number of very interesting and productive studies in recent years. A distinctive feature of those works is the consideration of a large number of different discursive factors that affect the choice of reference devices.

Of particular importance is the correlation of the processes of information extraction and replenishment of ontology. On the one hand, ontology is used to represent the results of information extraction, and on the other hand, the knowledge presented in the ontology helps solve specific problems of information extraction. The task of extracting information is considered as the task of identifying all references to objects of a given subject area (SA): entities, situations, events, states of objects, processes, etc. Found objects should be represented as instances of concepts and relations of SA ontology. Moreover, it is required to establish referential relationships between all objects found in the process of text analysis and instances of concepts and relations of the information content of ontology. This, in turn, does not exclude the possibility of adding new instances to the ontology.

This study uses a multifactorial quantitative approach to referencing, based on the notion that referential choice depends on the degree of activation of the referent in the focus of the speaker's attention [1]. The degree of activation, in turn, is associated with a number of factors that are determined by the properties of the referent, anaphora, or antecedent, and the structure of the text. To measure the level of activation within the framework of a given theory, the concept of "similarity measure" is used. The greater the value of the similarity measure, the greater the likelihood of using a reduced referential expression. The referential choice is not always a completely deterministic and categorical choice. For example, there are positions where only full name groups are used, and in some cases, only pronouns are used. There are intermediate cases where both full referential expressions and reduced ones can be used.

The measure of similarity used in a given cognitive multifactorial approach to measuring the degree of activation depends on a set of different factors. The weight of each factor is summed up to obtain a measure of similarity, and depending on its value, a reduced or complete referential agent is used. Activation factors affecting the choice of referential expression can be related both to the referent itself and to the context of the statement (Fig. 1). Among the discursive structure factors that influence referential choice are the different types of distances from anaphora to antecedent. Such as linear distance in clauses, sentences, paragraphs, and rhetorical distance. The rhetorical distance in the hierarchical structure of discourse presented within the framework of the rhetorical structure is understood as the distance from the anaphora to the antecedent.

The next important component of a given cognitive model is filters. In the process of resolving referential relations, there are cases of an equal measure of similarity in several referents at the same time – such a phenomenon suggests the emergence of a referential conflict (referential ambiguity). The reference conflict filter prohibits the use of a pronoun if there is an equal measure of similarity in more than one referent, thereby resolving the ambiguities that arise. A given filter does not affect the measure of referent similarity, so it is a separate component of the cognitive multifactor model. The tools that help resolve referential conflict are called referential deconflictors. Referential deconflictors in the cases where pronouns are used are grammatical gender, which, in turn, is consistent with the gender of only one of the referents. There are two types of deconflictors:

1. Conventional – based on the definition of grammatical gender (more suitable for the Russian-language and English-language texts).

2. Occasional – based on semantic compatibility with the context, knowledge about the world, and the exclusion of another referent from the number of possible due to its binding to another referential expression.

A referential conflict is considered resolved if reference deconflictors are triggered.

Modern advances in the field of computational linguistics are associated with the introduction of the latest methods of artificial intelligence based on neural networks, the development of methods for their integration with classical approaches, and the availability of a large amount of language data specially prepared to solve the problems of automatic text analysis. However, most of these resources are created for English, and there are very few corpora with reference markup. Accordingly, there is a need to devise a model for resolving referential relations in a multilingual system.

2. Literature review and problem statement

Paper [1] examines the referential choice between full name groups and anaphoric pronouns in the English-language newspaper articles. Machine learning methods are introduced using probabilistic characteristics, which are attributed by the logistic regression algorithm.

The authors of [2] consider referential conflicts in artistic poetic work. To this end, they used several discursive factors of referential choice. It is shown that with an active discursive factor, the status of the referent and its identification in the cognitive system of the speaker, the internal properties of the referent can resolve referential conflicts in the poetic lines. The intrinsic properties of the referent are understood as animacy, linear and rhetorical distances to the antecedent, discursive boundaries, and semantic-syntactic characteristics of the antecedent. All this suggests that it is advisable to conduct a study on the analysis of newspaper articles.

In work [3], the author establishes the specificity of the correlations of the antecedent and the referential means, considering them as the basis for the deployment of a complex discursive anaphora of a propositive nominal type, and also describes the specificity of two options for resolving the anaphora (directions of purpose and resources). The condition for modeling the linguistic situation of resolving the referent in the goal is the generation in the mind of the subject of a situation correlated with the target event, action, process, and reference content of the antecedent anaphoric complex. The structural minimum of the second direction of anaphora resolution is three-part structures.

The authors of [4] distinguish two types of factors that have a special impact on the assessment of the degree or measure of the referential proximity of two objects: discursive and semantic factors. Discursive factors are factors that are determined by the way objects in the text are expressed, their location relative to the structure of the text, as well as relative to each other. Semantic factors are factors that determine the assessment of the "similarity" of objects by their ontological structures and connections.

In work [5], the authors, in addition to the factors of the two types presented in [4], single out another factor – logical ontological. It makes it possible to consider the totality of the evaluation between objects and relies on the definition of the properties of relations given in the ontology. The proposed approach to the resolution of coreference is characterized by the following features:

– a distraction from discursive factors and emphasis on subject knowledge, primarily on the ontology of the subject area, regarding which the tasks of extracting information, removing ambiguity and resolving coreference are solved;

– scalability of the solution – the approach is applicable to the rules of information extraction and reference factors;

– independence (autonomy) – the approach is focused on fully automatic processing and does not require the input of the "correct" results of morphological analysis, the absence of grammatical errors, and complete parsing of sentences;

– integration of computational and linguistic models and methods of text analysis at the stage of semantic processing.

Thus, in order to resolve coreference, weighted coreferential connections between objects are established, while hypotheses (connections) are formed on the basis of a linguistic model, and resolution (choosing the best hypothesis) is formed on the basis of statistical data. The approach is tested on the texts of technical specifications from the subject area of automated control systems.

The approach proposed by the authors of [6] is based on subject knowledge, primarily on the ontology of the subject area. It provides extensibility with respect to information retrieval rules and referential factors. Also, the approach is focused on fully automatic processing and integrates computational and linguistic models and text analysis methods at the stage of semantic processing. A given approach is tested on the texts of technical specifications from the subject area of automated control systems.

The author of [7] attempts to consider the indexical shift from the standpoint of psycholinguistics of discourse. An indexic shift is understood as the phenomenon that some deictic expressions in additional subordinate clauses are interpreted not with respect to the context of the entire speech act but with respect to the coordinates given by the matrix clause. The factors affecting the referential choice are determined – distance, priming of one of the referents, and their mutual location. The experiment was conducted with speakers of the Mishar dialect of the Tatar language.

In work [8], the author comes to the conclusion that ambiguity in advertising is created intentionally. It is contextually conditioned and is usually resolved by means of pre-defined algorithms by the sender, which are "keys" in the form of information redundancy markers that orient the recipient to select one of the meanings. Such an algorithmic formation and resolution of ambiguity makes it possible to attract the attention of the recipient and creates the illusion of an independent resolution of the problem (removal of ambiguity). A given formation is a "programmed choice" in the actual absence of choice and corresponds to the linguistic manipulation characteristic of advertising discourse. The object of semantic modulations in the slogan – the classic advertising genre – becomes one linguistic unit. This contributes to the creation of semantic uncertainty, which, however, is easily resolved with the help of markers of information redundancy (indicating the name of the product, using the image of the product packaging on the poster, including a broad context). In the second stage of manipulation, the ambiguity is completely removed, leaving the addressee with a "choice" of a single programmed meaning.

The model reported in [9] makes it possible to obtain high-quality results for the recognition of coreferential bonds with an accuracy of up to 60 %. To improve the results, one can customize the model by changing (deleting/adding) the components of the parametric vector. At the same time, it is impossible to obtain an unambiguous universal model suitable for all types of coreference.

In [10], the author tries to define the concept of coreference, overviews existing approaches to the automatic resolution of coreference, substantiates the relevance of the study, and chooses an algorithm for solving this problem for the Russian language.

The above works [1–10] involved monolingual systems: English, Russian, and Tatar, and do not include the Kazakh language due to the lack of a marked multilingual corpus. The absence of such studies in the multilingual system (Kazakh, Russian, English) makes it possible to update the development of a model for resolving referential relations in this area. Therefore, devising a unique multilingual resource could help solve the problem associated with the need for automatic text processing in a multilingual system.

3. The aim and objectives of the study

The aim of this study is to devise a model for resolving referential relations in a multilingual system. This will provide an opportunity to build a multilingual resource to support national research in the field of computational linguistics and automatic text processing.

To accomplish the aim, the following objectives have been set:

– to build a multi-factor model for the selection of reference devices;

– to conduct an experimental study of the resolution of referential relations in a multilingual system.

4. The study materials and methods

To construct a model for resolving referential relations, the environment was considered. Ontology models a sig-

nificant part of the subject area for the user and provides structuring of information. We believe that the O ontology includes the set of classes C_0 describing the concepts of the subject area, the set of data domains D_0 (or data types), and the set of attributes $Atr_0=Dat_0\cup Rel_0$. Dat_0 are the attributes of simple data types, and R Rel_0 are the object attributes whose values are instances of classes from C_0. Each $c\in C_0$ class is defined by the set of attributes $c=(Dat_c,Rel_c)$. Each attribute of the simple type $\alpha\in Dat_c\subseteq Dat_0$ is mapped to a domain $d_\alpha\in D_O$ with values from the set of possible $V_{d\alpha}$ values. Each object attribute $\rho\in Rel_c\subseteq Rel_0$ takes values from the set of instances of classes $C_p\subseteq C_p$. The set of all attributes of a class is denoted as $Atr_c=Dat_c\cup Rel_c$. We consider ontology without synonyms of classes and data attributes, i. e. $\forall\ \alpha_1$, $\alpha_2\in Dat_O$:$d_\alpha\neq d_{\alpha 2}$ and $\forall c_1$, $c_2\in C_O$: $Atrc_1\neq Atrc_2$. The most important property of an ontology is the ability to define inheritance on classes: class c_2 inherits class c_1 ($c_1<c_2$), if and only $\forall a\in c_2$:$a\in c_1$. For the γ attribute, let's label its class as γ and its set of values as D_γ. Among the attributes of the class, we highlight a non-empty set of key attributes Atr_c^K, which ensure the identifiability (unambiguous definition) of class instances. Key attributes can be both simple type attributes and object attributes.

The set $a=(c_a, Dat_a, Rel_a)$ is an instance of the $c_a=(Dat_{ca}, Rel_{ca})$ ($a\in c_a$) class if and only each attribute of a simple type in Dat_a has the name $\alpha a\in Dat_{ca}$ with $V_{\alpha a}$ values from $V_{d\alpha a}$. Each relation attribute in Rel_a has the name $p_a\in Rel_{ca}$ with V_{pa} values as class instances from C_p. The IC_o content of an O ontology is a set of instances of ontology classes. The task of replenishing ontologies is to calculate the information content of the ontology from the input data.

Define the set A of information-text objects (i-objects) extracted from the input data and corresponding to the instances of the ontology classes. Each information object $a\in$ A has the form $(c_a, Dat_a, Rel_a, G_a, P_a)$, where $c_a\in C_o$ is an ontology class. Dat_a is the set of data attributes $\alpha_a=(\alpha, V_{\alpha a})$, where $\alpha\in Dat_{ca}$ is the attribute name, and $V_{\alpha a}$ is the set of $v\in d\alpha$ values; Rel_a is the set of object attributes $\rho_a=(\rho, V_{pa})$. $\rho\in Rel_{ca}$ is the name of the attribute, and is the set of i-objects of class $c_{pa}\in C_{pa}$; G_a is the grammatical characteristics. They are formed according to the grammatical characteristics of lexical objects, based on which a given i-object was obtained; P_a is the structural-textual information (the set of positions in the text and formal segments). The γ attribute of the i-object a is called populated if $V_{\gamma a}\neq\varnothing$. Let's denote the set of all the attributes of the i-object a as $Atr_a=Dat_a\cup Rel_a$. Each i-object naturally corresponds to some instance of an ontology: if $a=(c_a, Dat_a, Rel_a, G_a, P_a)$ is an i-object, then the corresponding instance of the ontology is $a'=(c_a, Dat_a, Rel_a)$. In this case, each $\alpha\in Dat_a$ attribute has values in $V_{\alpha a}$, and each $\rho\in Rel_a$ has values in V_{pa}.

The task of resolving the reference is to determine the correspondence of the data of i-objects (candidates for referents) to the same instance of the ontology.

Paper [4] considered three types of factors that affect the assessment of the degree or measure of referential proximity of two objects:

1. Discursive factors (local text and contextual) are determined by the way objects in the text are expressed, their location relative to the structure of the text and relative to each other.

2. Semantic factors determine the assessment of the similarity of objects by their ontological structure and connections.

3. Logical and ontological factors make it possible to consider the totality of relations between objects.

Within the framework of our work, it is possible to identify several more factors affecting the referential choice. The main task was not only to search for and study the factors affecting the referential choice but also to investigate their individual contribution to the accuracy of the prediction of referential choice, in order to reduce their number to the necessary minimum (to reduce the labor intensity of the annotation process). The full set of factors includes various characteristics of both the anaphora and the antecedent, as well as the referent itself, as well as some common discursive characteristics.

4. Attributes of the referent: animacy, gender, and number.

5. Attributes of the antecedent: whether the composition of direct speech is included, the type of syntactic group, the grammatical role, the referential form, the length of the antecedent in words, the number of antecedents in the chain from the current place to the full nominative group.

6. Attributes of an anaphora: the first/not the first mention in the discourse, whether it is part of direct speech, the type of syntactic group, the grammatical role, the number of references to the referent in the chain.

7. Distances between the anaphora and the antecedent: linear distance in words, linear distance in clauses, linear distance in sentences, distance in marcabules, the rhetorical distance of elementary discursive units, distance in paragraphs. Rhetorical distance, which is the length of the path between fragments of text along the constructed rhetorical network, is considered an important factor in referential selection. Rhetorical distance makes it possible to take into consideration the relationship between fragments of text that are far from each other in the linear distance but close in the structure of presentation.

During the study, several other factors influencing referential choice were added to the number of main factors.

The rhetorical distance to the antecedent f_1 defines the distance, which is measured in discourse units, from the current unit to the rhetorically closest containing the antecedent. Rhetorical distance is measured based on rhetorical structure. Rhetorical structure theory states that each unit of discourse is related to at least one other unit a certain "rhetorical relation," which is a sequence, cause, result, and so on. In the study, the theory of rhetorical structure is applied with a certain adaptation. In the rhetorical distance factor, a fourfold difference is presented: the rhetorical distance can be 1, 2, 3, and be greater than three. All kinds of references are understood as rhetorical antecedents in our work.

The syntactic and semantic role of the antecedent f_2 reflects the fact that those referents that were last referred to as subjects or actors in their clauses are more pronominalizable. Often, the properties of the subject and the actor coincide but the combinations of subject/non-actor and actor/non-subject are not excluded. We distinguish three situations: when the antecedent is both the subject and the actor; when it is either a subject or an actor; when it is neither.

The animacy f_3 is represented by two features: animacy and inanimateness. "Humanity" is an inherent property of the referent, which in some cases can increase the measure of its activation; the effect of animacy depends on rhetorical distance. At greater distances, "humanity" helps maintain activation at a higher level, and at shorter distances, inanimate referents receive and maintain activation to the same extent as "human" referents.

The linear distance f_4 may seem an optional parameter, as it is claimed that the rhetorical distance is the most powerful factor. However, a short rhetorical distance with a short linear distance is not the same as a short rhetorical distance with a long linear distance.

The spacing between the paragraphs f_5 reflects the importance of episodic structure in discourse. Usually, within a paragraph, activation is stored well while the paragraph boundary is reflected cognitively as an update to the activation distribution.

5. Results of studying the multifactor model of the choice of reference devices

5. 1. Multi-factor model of referral device selection

The counting system presented in [11] and its modification were selected by trial and error until the selected numerical weights began to explain all the available material. The model for calculating activation coefficients, devised for this study, is given in Table 1.

Table 1

Numerical weights of factor values

Attribute	Value	Weight
Rhetorical distance (RhD)	0; 1; 1.5	0.6
	2; 2.5; 3	0.5
	3.5	0.4
	≥4	0
Linear distance (LinD)	0	0.1
	1	0
	2	–0.1
	3	–0.2
	>3	–0.3
Animation	LinD≤2	0
	LinD≤3:Animate	0.2
	Inanimate	0.1
The syntactic role of the antecedent	RhD>3.5	0
	RhD≤3.5:Sudj	0.3
	Dir_Obj. Indir_Obj. Obl	0.2
	Attribute. Possessor	0.1
Distance between anaphor and antecedent in paragraphs (ParaD)	0	0
	1	–0.2
	>1	–0.4

Referential conflict is considered as a situation in which two non-referent *i*-objects are potential referents for some *i*-object. To determine which of these *i*-objects are true referents, we use the measure of similarity of *i*-objects. For *i*-objects *a* and *b*, let's denote this measure as $cs(a, b)$ (1). If non-referent *i*-objects *a* and *b* are candidates for referents for *i*-object *c*, then we consider that the reference conflict is resolved in favor of object *a* if and only $cs(a,c)>cs(b,c)$.

$$cs(a,b)=cs_{f1}(a,b)+cs_{f2}(a,b)+cs_{f3}(a,b)+ +cs_{f4}(a,b)+cs_{f5}(a,b). \quad (1)$$

Factors are used to assess the proximity, or similarity, of objects that are mentioned in the text. For each factor, an estimate of the distance $cs_f(a, b)$ is formulated, which reflects the degree or probability of a reference relationship between *i*-objects *a* and *b* as a function of factor *f*, without taking into consideration other factors.

Applying the counting model to the study material, the following correspondence was established between potential referential means and activation coefficients (Table 2).

Table 2

Correspondence between the measure of similarity and the type of referential expression

Preferential expression	Full noun phrase only (NP)	Full noun phrase pronoun	Full NP or pronoun	Pronoun full noun phrase	Pronoun only
Similarity measure	≤0.4	0.5	0.60.7	0.8	0.91

The data given in Table 2 demonstrate that the similarity measure values falling within the range of 0.5 to 0.8 can be called intermediate as they characterize cases of non-categorical referential selection.

Within the framework of the study, certain numerical values are assigned to the characteristics of the factors affecting the measure. All values of the activation factors are summed up so that the resulting activation score motivates the choice of a reference device. Table 1 gives five activation factors with their numerical values – a set of five factors is necessary and sufficient to explain the measure of the similarity of referents.

5. 2. Resolving referential relationships in a multilingual system

The procedure for calculating a measure of similarity is described as a method of prizes and penalties. Some factors increase, some decrease, and some have no effect on the numerical indicators measure of similarity.

Specific numerical values for each attribute were found empirically. For each mention of the referent, the current measure of similarity is calculated by adding the numerical values of the activation factors, both positive and negative. An excerpt of text in the Russian language is considered:

«СОФЬЯ КОВАЛЕВСКАЯ (1850–1891) – Первая в России женщина-профессор и первая в мире женщина-профессор математики. Открыла третий классический случай разрешимости задачи о вращении твердого тела вокруг неподвижной точки. Доказала существование аналитического решения задачи Коши для систем дифференциальных уравнений с частными производными, одна из теорем называется теоремой Коши – Ковалевской.»

Ковалевская:

<Ковалевская – (первая в России) женщина[1]> – nominal anaphora

<женщина[1] – профессор[1]> – associative anaphora

<Ковалевская – (первая в мире) женщина[2]> – nominal anaphora

<женщина[2] – профессор[2] (математики)> – associative anaphora

<Ковалевская – ∅[1]> – zero anaphora

Теорема:

<одна[1] (из теорем) – теорема Коши – Ковалевской> – cataphora

The text is divided into the following discursive units:

«0101 СОФЬЯ КОВАЛЕВСКАЯ (1850–1891)
0102 Первая в России женщина-
0103 профессор и
0104 первая в мире женщина-
0105 профессор математики
0106 ⦸Открыла третий классический случай разрешимости задачи о вращении твердого тела вокруг неподвижной точки
0107 ⦸Доказала существование аналитического решения задачи Коши для систем дифференциальных уравнений с частными производными
0108 одна из теорем называется теоремой Коши – Ковалевской.»

The «Ковалевская» referent considered in each line:
– line 0101 *a*=Ковалевская; *b*=Ковалевская (Table 3);
– line 0102 *a*=Ковалевская; *b*=женщина

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0103 *a*=Ковалевская; *b*=профессор

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0104 *a*=Ковалевская; *b*=профессор

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0105 *a*=Ковалевская; *b*=женщина

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0106 *a*=Ковалевская; *b*=

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0107 *a*=Ковалевская; *b*=

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0108 *a*=Ковалевская; *b*=теорема

cs(*a*,*b*)=0.6+0+0+0+0=0.6.

Table 3

Example of calculating a similarity measure

RhD	1	0.6
LinD	1	0.1
ParaD	0	0
Animation	no, LinD≤2	0
Synt. and semantic. role	Active S LinD≤2	0.3
cs(*a*,*b*)	–	1

The «теорема» referent considered in each line:
– line 0101 *a*=теорема; *b*=Ковалевская (Table 4);
– line 0102 *a*=теорема; *b*=женщина

cs(*a*,*b*)=0+(–0.3)+0+0+0=–0.3;

– line 0103 *a*=теорема; *b*=профессор

cs(*a*,*b*)=0+(–0.3)+0+0+0=–0.3;

– line 0104 *a*=теорема; *b*=профессор

cs(*a*,*b*)=0+(–0.3)+0+0+0=–0.3;

– line 0105 *a*=теорема; *b*=женщина

cs(*a*,*b*)=0+(–0.3)+0+0+0=–0.3;

– line 0106 *a*=теорема; *b*=⦸

cs(*a*,*b*)=0.4+(–0.2)+0+0+0=0.2;

– line 0107 *a*=теорема; *b*=⦸

cs(*a*,*b*)=0.5+(–0.1)+0+0+0=0.4;

– line 0108 *a*=теорема; *b*=теорема

cs(*a*,*b*)=0.6+0.1+0.3+0+0=1.

Table 4

Example of calculating a similarity measure

RhD	RhD≥4	0
LinD	LinD≥3	–0.3
ParaD	0	0
Animation	no. LinD≤2	0
Synt. and semantic. role	S LinD≤2	0
cs(*a*,*b*)		–0.3

An excerpt of the text in English is considered:

«*SOFIA KOVALEVSKAYA* (1850–1891) – *Russia's first female professor and*
the world's first female professor of mathematics. body around a fixed point. She proved the existence of an analytical solution of the Cauchy problem for systems of differential equations with partial derivatives, one of the theorems is called the Cauchy – Kovalevskaya theorem.»

The "*Kovalevskaya*" referent is considered in each line:
– line 0101 *a*=«*Kovalevskaya*»; *b*=«*Kovalevskaya*» (Table 5);
– line 0102 *a*=«*Kovalevskaya*»; *b*=«*female*»

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0103 *a*=«*Kovalevskaya*»; *b*=«*professor*»

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0104 *a*=«*Kovalevskaya*»; *b*=«*professor*»

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0105 *a*=«*Kovalevskaya*»; *b*=«*female*»

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0106 *a*=«*Kovalevskaya*»; *b*=∅

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0107 *a*=«*Kovalevskaya*»; *b*=∅

cs(*a*,*b*)=0.6+0+0+0+0.3=0.9;

– line 0108 *a*=«*Kovalevskaya*»; *b*=«*theorem*»

cs(*a*,*b*)=0.6+0+0+0+0=0.6.

Table 5

Example of calculating a similarity measure

RhD	1	0.6
LinD	1	0.1
ParaD	0	0
Animation	no. LinD≤2	0
Synt. and semantic. role	Active S LinD≤2	0.3
cs(*a*,*b*)		1

The "*theorem*" referent is considered in each line:
– line 0101 a=«*theorem*»; *b*=«*Kovalevskaya*» (Table 6);
– line 0102 «*theorem*»; *b*=«*female*»

cs(*a*,*b*)=0+(–0.3)+0+0+0=–0.3;

– line 0103 «*theorem*»; *b*=«*professor*»

cs(*a*,*b*)=0+(–0.3)+0+0+0=–0.3;

– line 0104 «*theorem*»; *b*=«*professor*»

cs(*a*,*b*)=0+(–0.3)+0+0+0=–0.3;

– line 0105 «*theorem*»; *b*=«*female*»

cs(*a*,*b*)=0+(–0.3)+0+0+0=–0.3;

– line 0106 «*theorem*»; *b*=∅

cs(*a*,*b*)=0.4+(–0.2)+0+0+0=0.2;

– line 0107 «*theorem*»; *b*=∅

cs(*a*,*b*)=0.5+(–0.1)+0+0+0=0.4;

– line 0108 «*theorem*»; *b*=«*theorem*»

cs(*a*,*b*)=0.6+0.1+0.3+0+0=1.

Table 6

Example of calculating a similarity measure

RhD	RhD≥4	0
LinD	LinD≥3	–0.3
ParaD	0	0
Animation	no. LinD≤2	0
Synt. and semantic. role	S LinD≤2	0
cs(*a*,*b*)		–0.3

An excerpt of the text in the Kazakh and Russian languages is considered:

«*Toyota Motor* распространила сообщение о планах производства на апрель, а также напомнила о временной остановке предприятий в Японии. Заводы марки перестанут работать 22 марта. Простой продлится 8 дней. Причиной является нехватка электронных компонентов.»

Toyota Motor – заводы:

0101 Toyota Motor распространила сообщение о планах производства на апрель,
0102 а также Ø напомнила о временной остановке предприятий в Японии.
0103 Заводы марки перестанут работать 22 марта.
0104 Простой продлится 8 дней.
0105 Причиной является нехватка электронных компонентов.

Check the line 0102 *a*=Toyota Motor; *b*=Ø (Table 7).

Table 7

Example of calculating a similarity measure

RhD	1	0.6
LinD	1	0
ParaD	0	0
Animation	no, LinD≤2	0
Synt. and semantic. role	Active S LinD≤2	0.4
cs(*a*,*b*)		1

An example of the text in the Kazakh language

«Сәуірде қанша көлік өндіруге ниетті екенін хабарлаған Toyota Motor Жапониядағы кәсіпорын 22 наурызда жабылып, 8 күн жұмыс істемейтінін де ескертті. Бұған электронды компонент тапшылығы себеп болған.»

Similar to the previous examples, it is broken down into discursive units:

«0101 Сәуірде қанша көлік өндіруге ниетті екенін хабарлаған Toyota Motor
0102 Жапониядағы кәсіпорын 22 наурызда жабылып,
0103 Ø 8 күн жұмыс істемейтінін де ескертті.
0104 Бұған электронды компонент тапшылығы себеп болған.»

Checking the line 0102 *a*=«Toyota Motor»; *b*=«кәсіпорын» (Table 8).

Table 8

Example of calculating a similarity measure

RhD	1	0.6
LinD	1	0
ParaD	0	0
Animation	no, LinD≤2	0
Synt. and semantic. role	Active S LinD≤2	0.4
cs(*a*,*b*)		1

According to the simulation data, the accuracy of the algorithm was 88 %, among all forms, 55 were predicted incorrectly. All deviations can be divided into two categories. The

first type – the referential expressions can be selected with approximately equal probabilities – of the variant falls in the range from 0.5 to 0.55. The second type is those predictions in which the difference between probabilities varies from 0.1 to 0.8. Table 9 illustrates the types of deviations.

Table 9

Example of algorithm deviations

Deviation type	Reference	Predicted form	Reference probability	Predicted form probability
I	Pronoun	Full Noun Phrase	0.44	0.56
II	Full Noun Phrase	Pronoun	0.394	0.606

It is possible to focus on the speaker's message through grammatical roles. The focus of attention in the three languages is successively encoded by the speaker as the subject of the clause. Subjectivity and reduced forms of reference are in a causal relationship: antecedent subjectivity is one of the most powerful factors leading to the choice of a reduced form of reference. In both English and Russian, Kazakh languages, antecedent subjectivity can add points to the general measure of similarity of the referent. In both English and Russian and Kazakh discourses, 86 % of pronouns that do not allow a reference alternative have a subject as an antecedent.

6. Discussion of results of studying the multi-factor model of selecting reference devices

During the study, all types of referential relations were analyzed and a theoretical study of methods for their solution was carried out. Works [1–10] consider the referential choice only between full name groups and anaphoric pronouns, in addition, studies were conducted on monolingual systems (the English, Russian, Tatar languages). In our study, an attempt is made to integrate the model from [11] into the proposed approach [4, 5]. A model for resolving referential relations in a multilingual system is proposed. This makes it possible to design a multilingual resource to support national research in the field of computational linguistics and automatic text processing.

The proposed model is an extension of the rules for extracting information and referential factors in a multilingual system. The peculiarity of the model is in the integration of computational and linguistic models and methods of text analysis at the stage of semantic processing. The features also include the presence of the Kazakh language in this system. This results in faster automatic word processing. A multifactorial model for the selection of reference devices is proposed (Table 1).

An experimental study of the resolution of referential relations in a multilingual system was carried out. For the examples, proximity measures were calculated manually (Tables 3–8). The strongest relationship between the measure of proximity and the probabilistic characteristic is observed in the case of correct predictions. It was also estimated that 75 % of all deviations had a similarity measure of 0.5 to 0.8. Consequently, the non-categorical nature of referential selection is one of the main causes of deviations in the modeling of referential expressions.

The biggest limitation in the use of this model is the lack of a marked multilingual case. Advancing this study could lead to the development of an intelligent information resource on modern methods of automatic text processing. The information resource to be designed would provide convenient meaningful access to information about a given method of automatic text processing, pre-trained models, test data, marked text corpora, and other information resources on this topic.

7. Conclusions

1. A model for the selection of reference devices has been proposed, consisting of a set of factors that can either increase or decrease the measure of similarity of a particular referent. The features of each factor have certain numerical values, respectively, for each referent, a measure of the similarity of the cognitive and at the same time the numerical equivalent of pronominalizability is calculated. Among the factors that are crucial for the multilingual system sample are the rhetorical distance to the antecedent, the syntactic and semantic role of the antecedent, the animation, the linear distance to the antecedent, and the paragraph distance of the referents.

2. The proposed model for the selection of reference devices, designed to calculate the measure of proximity of two objects, has been tested in a multilingual system. The effectiveness of the model is evaluated using such a metric as accuracy. This metric is the ratio of correctly predicted forms to the total number of predicted referential expressions. According to the data from our simulation, the accuracy of work was 88 %, among all forms.

Acknowledgments

"This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No AP09057872)"

References

1. Kudriavtceva, A. S. (2020). Referent activation and probabilistic evaluation of referential choice: a study of English newspaper texts. Computational Linguistics and Intellectual Technologies.
2. Zhanturina, B. N., Makarenko, A. S. (2021). Referential ambiguity and discourse factors. Voenno-filologicheskiy zhurnal, 3, 13–21.
3. Voronina, L. V. (2020). Relevance of the reference and referential mean within the antecedent-anaphoric complex with purpose semantics in political discourse. Bulletin of the Moscow State Regional University (Russian Philology), 5, 16–25. doi: https://doi.org/10.18384/2310-7278-2020-5-16-25

4. Garanina, N. O., Sidorova, E. A., Seryi, A. S. (2018). Multiagent Approach to Coreference Resolution Based on the Multifactor Similarity in Ontology Population. Programming and Computer Software, 44 (1), 23–34. doi: https://doi.org/10.1134/s0361768818010036
5. Sidorova, E. A., Garanina, N. O., Kononenko, I. S. (2018). Mnogomestnye ontologicheskie otnosheniya v zadache razresheniya koreferentsii. Shestnadtsataya natsional'naya konferentsiya po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2018.
6. Sidorova, E. A., Garanina, N. O., Kononenko, I. S., Sery, A. S. (2018). Approach to coreference resolution based on ontological similarity measure. Intellekt. Yazyk. Komp'yuter, 1, 347–351.
7. Ganieva, S. K. (2021). Indexical shift: typology and analysis. Aktual'nye problemy yazykoznaniya, 49–56.
8. Sokolova, O. V. (2021). Lingvopragmaticheskie i semanticheskie parametry yazykovoy i diskursivnoy kreativnosti v reklame. Kritika i semiotika, 2, 52–70.
9. Solov'ev, S. S., Garshina, V. V. (2020). Ispol'zovanie mashinnogo obucheniya dlya razresheniya koreferentsii. Sbornik studencheskikh nauchnykh rabot fakul'teta komp'yuternykh nauk VGU, 259–265.
10. Kupriyanova, A. D., Shilin, I. A. (2018). Primenenie metodov mashinnogo obucheniya k zadache razresheniya koreferentsii. Al'manakh nauchnykh rabot molodykh uchenykh Universiteta ITMO, 2, 387–389.
11. Kibrik, A. A. (1999). Reference and Working Memory. Current Issues in Linguistic Theory, 29. doi: https://doi.org/10.1075/cilt.176.04kib