International Journal of Computer Applications Technology and Research
Volume 12–Issue 04, 39-42, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1204.1009

An Algorithm for Finding Equivalence in Referential Relations

Kalman Gulzhamal
doctoral student of specialty "Information System" at L.N.Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan

Ilyubayev Adelzhan
Teacher,Master of Technical Science, Abay Myrzakhmetov Kokshetau University, Kokshetau, Kazakhstan

Aktaeva Dilara Aidosovna
Teacher,Master of Technical Science, Abay Myrzakhmetov Kokshetau University, Kokshetau, Kazakhstan,

Kasym Karlygash Kydyrbekkyzy
Teacher, Master of Technical Science, Abay Myrzakhmetov Kokshetau University, Kokshetau, Kazakhstan

Likerova Zinaida Valentinovna
Teacher, Master of Technical Science, Abay Myrzakhmetov Kokshetau University, Kokshetau, Kazakhstan

Dauletbek Aigerim
Graduate Student of Specialty "Information System" at Abay Myrzakhmetov Kokshetau University, Kokshetau, Kazakhstan,

Abstract: Establishing referential relations in discourse is one of the most relevant but difficult to model problems of automatic text analysis. Reference means attributing a textual unit (linguistic expression) to a non-linguistic object (referent). The correct interpretation of the statement in the analyzed text requires solving the referent of the textual reference about the object, that is, the reference of the textual expression. In this work, a model for resolving referential relations is proposed.

The method used to solve the referential relationship is finding equivalence, that is, by calculating the degree of similarity of possible elements in the word, the similarity in the referential relationship is calculated, and as a result, the relationship value of the referent in the sequence of sentences is obtained.

These opportunities will be a new beginning for solving referential relationships in the Kazakh language. Based on the possibilities of computer technology, this direction in developing language education creates a great opportunity for extensive research and analysis at various levels of the language treasure as a bright mirror of the spiritual and material culture of the people.

Keywords: Anaphora, coreference, discourse, information object, ontology.

1. INTRODUCTION

Anaphoric, cataphoric, and coreference are the main elements of resolving referential relations.

Anaphora is a relation between sentences in which the meaning of one word or phrase is defined by another word or phrase. The first member of an anaphoric relation is called an antecedent, and the second member is called an anaphor or anaphor.

A cataphoric relation is contrasted with an anaphora where the anaphora occurs in the first clause and the antecedent occurs in the second clause. Anaphoric and cataphoric relations are formed from nouns.

The most common form of anaphoric relation is pronoun anaphora. This type of anaphora includes the third person form of the pronoun, among the pronouns, the most anaphoric functions are the classification pronoun and the reference pronoun. We can see the anaphoric function of the relative pronoun in the following examples

For example: *Труба түбіндегі жапырық тас үй – **мехцех**. **Бұл** - әншейін келешегіне қарай қойылған ат, әйтпесе нобайы түзу бір механизм жоқ.(The leafy stone house at the bottom of the pipe is a mehtzeh. This is a name given to the future, otherwise there is no straight mechanism.)*

This syntactically complex unit that connects the first sentence and the second sentence is an anaphoric relationship, where the word **мехцех** in the first sentence is repeated with pronoun **бұл** in the second sentence.

***Елжас** өткен айда Бразилияда болды. **Ол** сол елден сағаң сыйлық алып келінті. (Eljas was in Brazil last month. He brought you a gift from that country)*

In the example of the first sentence, it is anaphoric to have the classificatory pronoun **Елжас** and in the second sentence the word **ол** is repeated through it in third person form and repeated with the same indicative pronoun.

Solving this problem faced by natural language is the most difficult for other languages and has not yet been fully resolved. For example, the scheme described in [6] includes mentions of people and organizations and the task of solving anaphora, while the task of solving anaphora in [7] and [8] is limited to pronoun anaphora.

International Journal of Computer Applications Technology and Research
Volume 12–Issue 04, 39-42, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1204.1009

Solving anaphora is a very important task, many researchers take different approaches to this problem and use different methods: traditional (syntactic and semantic) and alternative (statistical) methods, recently corpus and ontological methods are also used.

The corpus method is a collection of texts within the discipline and thus research.

Most of the similar resources available today are English-language resources. A research group from Stanford University recommends using Wikipedia to solve co-references [4]. The approach itself is based on the use of several simple filters together. The system based on this approach has now been expanded with new filters [2]. Two of the five proposed new filters use external resources such as WordNet [8], Wikipedia, and Freebase [5]. Projects such as WordNet and Freebase are suitable for the English language, which has a significant impact on research in the field of English text processing.

Solving such relations for the Kazakh language will be very complicated, however, in solving the referential relations for the Kazakh language, considering the semantic analysis model and the peculiarities of the Kazakh language, combining similar elements in the sentence, searching for corresponding elements, etc. a model for solving referential relations in the Kazakh language is created by using methods.

2. METHOD

2.1 The Architecture of Anaphora Resolution

In this method, we used the primary analysis component and the semantic analysis model, identification component, and data validation component.

This section presents the architecture of the anaphora resolution system in the Kazakh language (figure 1).

Fig.1. Reference resolution model.

The general operation of the model is as follows:

- Primary analysis. The information objects obtained from the document are sent to the primary analysis component, where checks are made for the presence of correlation and the matching of the main attributes with the DB objects according to the tuple. Information objects that have achieved correspondence with a single object of the database or a set of basic attributes are fully defined are identified.
- Identification component. The rest of the information objects fall into the identification component, where, if necessary, the collections of the closest objects of the database are filtered, which are expanded in the hierarchy of ontology classes or in other relationships of the ontology.
- Semantic analysis model. It helps to fully reveal the semantic relationship of the elements of the word, the meaning of the word, the semantic analysis model allows to speed up the work of the remaining components.

In the Kazakh language, the order of the words in a sentence is, in general, stable, not as free as in the Russian language, as a rule, the narrative is at the end of the sentence, the initial is before it, the determiner, the modifier, the complement is before the related words.

Pronouns often have the primary, complementary function in a sentence.

In this study, we take data sets from tengrinews news collection and stories of G. Mustafin: personal pronouns, demonstrative pronouns, reflexive pronouns are resolved in this dataset. A proper anaphora resolution system requires subject object matching. The proposed training dataset is defined as a subject, object, number, animate or inanimate object. This POS tagging system for Part-of-Speech Tagger for Kazakh language we use annotated corpus, Noun and Noun Phrase are selected using rules for Kazakh text.

2.2 Equivalence calculation algorithm

When searching for similar objects, candidate attributes are searched by pairwise comparison of the set of their values. To calculate the D-similarity measure, we obtained the following values.

$$a, b \in A, a \approx b, and\ e\ \alpha_a \in Dat_a, \beta_b \in Dat_b, \rho_a \in Rel_a, \vartheta_b \in Rel_b$$

The similarity measure for data attributes is determined by the number of equal values of the attributes:

α_a similarity across data

$$\beta_b(\alpha_a \sim \beta_b) \Leftrightarrow (\alpha = \beta \vee \alpha \ll \beta \vee \alpha \gg \beta) \forall N^d and\ S^d = V_{a_a} \cap V_{\beta_a} \neq \emptyset$$

D is similar in terms of data:

$$D(\alpha_a, \beta_b) = \frac{|S^d|}{2}\left(\frac{1}{|V_{a_a}|} + \frac{1}{|V_{\beta_a}|}\right) \quad (1)$$

Where $\mathrm{Dat_a}$ set of data attributes, $\mathrm{V_{a_a}}$ a set of information values, $\mathrm{Rel_a}$ a set of relationship attributes, $\mathrm{S^d}$ structured information (input data), α_a, β_b —co-referent-candidate.

The algorithm for finding coreference is as follows:

a and b are candidate coreferents

$$a \approx b \Leftrightarrow c_a = c_b \vee c_a < c_b \vee c_a > c_b\ and\ Atr_a^k \subseteq Atr_b^k \vee Atr_b^k \subseteq Atr_a^k$$

$$coR(x) = \cup_{x \in X} coR(x) \quad (2)$$

International Journal of Computer Applications Technology and Research
Volume 12–Issue 04, 39-42, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1204.1009

The degree of similarity of possible candidates is calculated from the formula (1) given above, and the algorithms for calculating the coreference using the co-referent-candidate *a* and *b* given by the formula (2) are given.

The model and calculation algorithms we used in this work are not designed to solve all types of referential relations, these calculation algorithms may not work properly due to the influence of factors affecting anaphora and cataphora and their types in sentences.

2.3 Combining information objects

D is a binary relation in the set of information objects, s^1 *and* s^2 objects, *and* $s^1 D s^2 \Leftrightarrow s^1 and s^2$ recognized referentially identical. Relationship D is the ratio of equivalence and, therefore, it divides many O_Q into non-overstepping subset of the clusters. Equivalence classes in relation to D coincide with components of the coherence of the column in Fig. 2. All information contained in the elements of the cluster is combined in one object called a nodal object or not. The g-equivalent of the chain of objects is considered nodular. Obviously, there will always be such (Fig. 3).

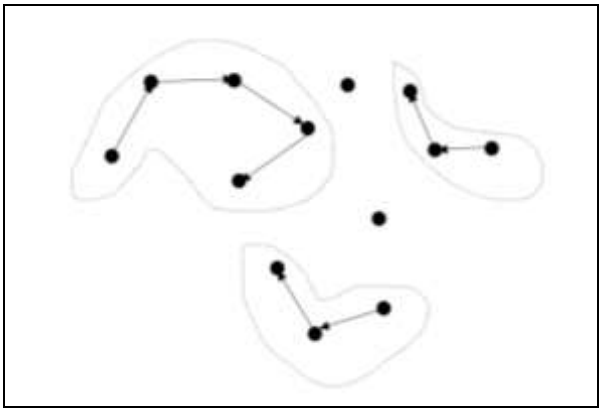

Fig.2. Marking up multiple objects

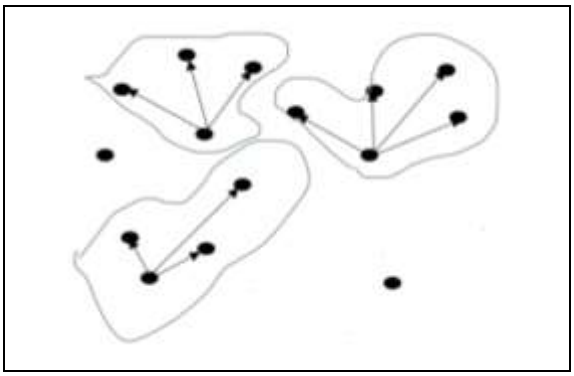

Fig. 3. Node objects

Combining objects in accordance with relationships, shown in fig. 3, we achieve that the cardinality of the set of objects becomes equal to cardinalities of the set of equivalence classes with respect to relation D. It's clear, what. $|O_Q/D| \leq |O_Q|$

3. EXPERIMENTAL PROCEDURE AND RESEARCH RESULT

To realize the algorithm, we used the ontology created in the subject area of computer sciences.

For practical tests of our method and obtaining the first approximate values of f *and* k, an editor of objects with built -in mechanisms for calculating the similarity of any two selected s^1 *and* s^2 objects and resolving reference connections in each many information objects was developed.

The selected fragment, 300 words, contain all referentially identical objects. Volume the total text is 420 words, the number of extracted objects is 30. Objects that are not included in this fragment do not fundamentally affect the result of resolving referential links. Ontological class instances have been retrieved: The result of participate (қатысу), Internet resource (интернет ресурстар), Scientific activity (Қызмет), Organization (ұйым) and Person(тұлға).

The result of scientific activity (Қызмет), class has 4 attributes: title, number, start date, completion date.

of Organization (ұйым) class has 4 attributes: e mail, abbreviation, address, founding date, organization name, organization description, phone number.

of participate (қатысу) class has 5 attributes: involved end, involved start, involved role, person involved, involved activities.

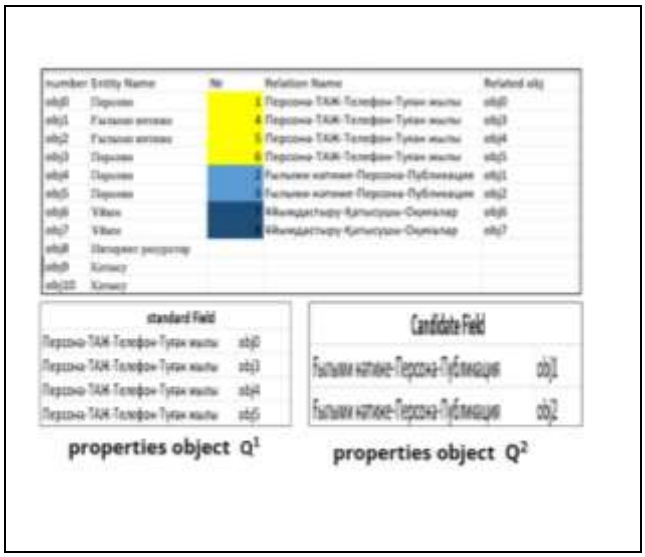

Fig. 4. Main window of the object editor

as you can see from the picture, no attributes of q^2 have been determined, we look for a candidate close to q^2 based on the above equivalence search algorithm, Obviously, since both objects are instances of the Scientific Event class, therefore the object q^1 is the equivalent of the object q^2.

Candidate(Q^1Q^2)=1

As a result of comparing the class attributes of the ontology, it is possible to find the equivalence of the attributes, for the purposes of our research, this is called coreference, that is, the coreference of the attributes with each other.

The general results of the research work can be seen in the table below.

Table 1. the result of finding the referent

Amount of text	Number of ontological classes	Number of attributes	Referent number of attributes
300	87	156	64
245	75	123	48
345	57	103	56
Total	219	382	168

4. CONCLUSIONS

In this article, we proposed a method for solving refractive relations, considering as an example the main elements of referential relations and their differences from each other, this method was obtained based on the algorithm for calculating the degree of similarity of attributes and semantic analysis, the algorithm for solving the referential relationship based on determining the mutual semantic relationship of elements in a word and finding the degree of similarity of a potential candidate, this work plays a key role for us in the development of work on solving the refractive relationship in the kazakh language.

REFERENCES

1. Caroline V. Gasperin Statistical anaphora resolution in biomedical texts. Technical report, University of Cambridge Computer Laboratory. 2009. ISSN 1476–2986
2. Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL2011 Shared Task. In Proceedings of the CoNLL2011 Shared Task.
3. Mitkov, R. Anaphora resolution: the state of the art, Working paper, (Based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolver hampton, 1999.
4. Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Sur deanu, Dan Jurafsky, and Christopher Manning. 2010. A Multi-Pass Sieve for Coreference Resolution. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)
5. The main page of the Freebase project. [Electronic resource] - Access mode http://www.freebase.com/
6. Ermakov A.E. Reference designations of persons and organizations in Russian-language media texts: empirical patterns for computer analysis. // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2005". M.: Nauka, 2005.
7. Zagorulko Yu.A., Borovikova O.I., KononenkoI.S., Sidorova E.A. An approach to building a subject ontology for a knowledge portal in computational linguistics. // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2006". M.: RGGU, 2006.
8. Official website of Princeton University. [Electronic resource]. Main page of the WordNet project. – Access mode http://wordnet.princeton.edu
9. Potepnya V.N. Resolution of pronominal anaphora in multilingual information systems. // Artificial intelligence-2006 №4 P.619-626.
10. [10] Gray A.S., Sidorova E.A. Identification of objects in the task of automatic document processing. // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2011". M.: RGGU, 2011. S. 580-591.
11. Tolpegin P.V., Vetrov D.P., Kropotov D.A. Algorithm for automated resolution of third person pronoun anaphora based on machine learning methods. Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2006" // Ed. N.I. Laufer, A.S. Narinyani, V.P. Selegeya. - M.: RGGU, 2006