Lecture Notes in Networks and Systems 1177

Atulya Nagar
Dharm Singh Jat
Durgesh Mishra
Amit Joshi *Editors*

Intelligent Sustainable Systems

Selected Papers of WorldS4 2024, Volume 3

Springer

Lecture Notes in Networks and Systems

Volume 1177

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Atulya Nagar · Dharm Singh Jat · Durgesh Mishra · Amit Joshi
Editors

Intelligent Sustainable Systems

Selected Papers of WorldS4 2024, Volume 3

Springer

Editors
Atulya Nagar
School of Mathematics, Computer Science and Engineering
Liverpool Hope University
Liverpool, UK

Dharm Singh Jat
Department of Computer Science
Namibia University of Science and Technology
Windhoek, Namibia

Durgesh Mishra
Symbiosis University of Applied Sciences
Indore, Madhya Pradesh, India

Amit Joshi
Global Knowledge Research Foundation
Ahmedabad, Gujarat, India

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-981-97-8694-7 ISBN 978-981-97-8695-4 (eBook)
https://doi.org/10.1007/978-981-97-8695-4

The Problem of Coreference in NLP

Utlir Khamdamov, Elov Botir, Axmedova Xolisxon, Abdisalomova Shahlo, Rakhimova Bekposhsha, and Hamroqulova Marjona

Abstract Artificial intelligence methods for performing a query in search engines require effective solution of the task of automatic processing of texts in the database. One of the important tasks of implementing machine learning is the development of a rule system. This article discusses the problem of coreference calculated from preliminary studies aimed at automatic determination of meaningful distances expressed in Uzbek language texts. Basic rules and stages of solving the coreference problem are also analyzed. Explanations of a number of terms such as anaphoric relations and their main types, the main used anaphora, antecedent, cataphora, clustering, singleton, etc., which create the coreference problem, are given.

Keywords NLP · Coreference resolution · Mention detection · NER · Pipeline conveyor · Tokenization · Coreference · Proper nouns · Pronoun recognition

U. Khamdamov
Department of Communication and Digital Technologies, Tashkent, Uzbekistan

E. Botir (✉) · A. Xolisxon · A. Shahlo · H. Marjona
Tashkent State University of Uzbek Language and Literature Named After Alisher Navoi, Tashkent, Uzbekistan
e-mail: elov@navoiy-uni.uz

A. Xolisxon
e-mail: xolisa9029@mail.r

A. Shahlo
e-mail: abdisalomovashahlo@navoiy-uni.uz

R. Bekposhsha
Urgench State University, Urgench, Uzbekistan
e-mail: filolograhimova726@gmail.com

A. Nagar et al. (eds.), *Intelligent Sustainable Systems*, Lecture Notes in Networks and Systems 1177, https://doi.org/10.1007/978-981-97-8695-4_14

1 Introduction

In this article, we will consider the issue of coreference in Uzbek language texts. It is one of the tasks belonging to the speech analysis part of NLP, and the sentences in the text form a single research object. Solving the issue of coreference in the text is a clustering task of NLP, which serves to determine the content distance between all names referring to one object/event in a sentence/document/corpus (all units naming one referent) [1, 2].

As an example, language units that represent an organization, person, place, object name can be given. Many natural language applications such as question–answer systems in NLP, automatic summarization of documents or machine translation need to identify references to objects in the text at the initial step [3–5]. To understand the coreference issue, let's consider the following example:

Nilufarning aytishicha, ***Azimjon*** *juda iltifotli yigit, chunki* ***u*** *har doim* ***Aziza****ni ishga* ***o 'zi****olib keladi, mashinaning eshigini ham* ***o 'zi*** *ochib,* ***ayolini*** *ishxonasiga kuzatib qo'yadi. (According to Nilufar, Azimjon, is a very kind man, because he always brings Aziza to work, opens the door of the car himself and escorts her to his office.)*

We highlight all notes in this sample that can name a subject with a single referent in a different color. These are:

Azimjon, U, O'zi, Aziza, Ayol

Therefore, the automatic determination of exactly which of these indicated notes have the same referent represents the essence of solving the coreference problem.

The task of automatic detection of coreference problems in the text can be carried out in the following two steps:

1. Identify object reference names; (object-reminiscent names-note)
2. Clustering.

1. Determining Names (Notes) in Relation to the Object

   The main goal of this step is to identify all the candidates that can refer to the objects. For example, the following example highlights candidate applications:

*Nilufar*ning Aytishicha, *Azimjon* Juda Iltifotli *Yigit*, Chunki *U* Har Doim *Aziza*ni Ishga *O'zi*olib Keladi, Mashinaning Eshigini Ham *O'zi*ochib, *Ayolini* Ishxonasiga Kuzatib Qo'yadi.

At this stage, three different links are identified [6–8]:

– pronouns are defined;
– nouns (NER) are defined;
– noun phrases (NPs) are defined.

1.1 Identifying Pronouns

Pronouns are words with an independent meaning—nouns, adjectives, numbers, adverbs, etc., which do not indicate a person, object, character or quantity, but are used instead of words, phrases and sentences, and refer to them. Since pronouns have a vague and general meaning, their main meaning and which word group they are used for is determined according to the text content. This semantic feature of pronouns can be seen in the following text:

Zamira *bugun uyda.* ***U*** *hovli-joylarni supurib, kir yuvishga tutindi.* ***Guldor*** *kiyimlarni alohida ajratdi, chunki* ***bunday*** *kiyimlarni boshqalariga aralashtirib bo 'lmaydi.* ***Bitta*** *sovunni maydalab kir mashinasiga solar ekan, "Oyim ham* ***shuncha*** *sovun ishlatarmidilar",—deb o'yladi. (Zamira is at home today. She swept the yards and did the laundry. The florist separated the clothes separately, because such clothes cannot be mixed with others. As he crushed one bar of soap and put it in the washing machine, he thought, "My mother would use so much soap.").*

In this text, pronouns *u, bunday, shuncha* are used instead of Zamira (noun), guldor (adjective), bitta (number), and they perform the same function as they did in the sentence. However, if these sentences are separated from the text, their meaning becomes abstract. Pronouns are divided into 7 different types according to their relative meanings and grammatical signs [9]:

1. *Personal pronouns* are used instead of personal nouns and refer to them.
2. *Demonstrative pronouns*—Person, object, sign, action expressed through independent words.
3. *Interrogative pronouns*—pronouns used to ask about a person, thing, sign-property, action-state.
4. *Disjunctive pronouns* are pronouns that take the place of independent words and summarize their meaning.
5. *Infinitive pronoun*—come in the place of independent words denoting person-thing, sign-property, and indicate its negation.
6. *Reflexive pronouns*—is used instead of personal pronouns in all three persons and emphasizes its meaning.
7. *Indefinite pronoun*—pronouns that take the place of independent words denoting a person-thing, sign-property, guess its meaning and suspect it.

Types of pronouns are listed in Table 1.

A pronoun is a word that replaces nouns and noun compounds and usually includes anaphora, in which the meaning of the pronoun depends on the preceding sentence (antecedent). But not all types of pronouns mentioned above form anaphoric relations. Some pronouns (3rd person personal pronouns, demonstrative pronouns, reflexive pronouns) represent anaphora in linguistics and depend on an antecedent. For example:

A 'zam *bilan Toshkentda ko'rishgan edim, bu xabarni o 'sha aytib berdi.*
(I Met Azam in Tashkent, He Told Me This News.)

In this sentence, the personal pronoun "u" and the word "A 'zam" have one referent, and the pronoun refers to the noun in the previous part of the sentence.

Table 1 Uzbek pronouns

Personal pronoun	Demons-trative pronouns	Interro-gative pronouns	Disjunc-tive pronouns	Infinitive pronoun	Reflexive pronouns	Indefinite pronoun
Men	U	kim?	hamma	hech kim	o ‘z	Kimdir
Sen	Bu	nima?	barcha	hech nima		Nimadir
U	Shu	qaysi?	bari	hech qancha		Qaysidir
Biz	o ‘sha	qancha?	har kim	har qanday		Allakim
Siz	mana bu (manavi)	necha?	har nima	har qaysi		Allanarsa
Ular	mana shu	nimaga?	Har qancha	hech bir		Allaqanday
	Ushbu	qani?	har bir			

Table 2 Pronouns representing anaphora in the Uzbek language

Personal pronouns	Demonstrative pronouns	Possesive pronouns
	U	*o ‘z*
	o ‘sha	
u, ular	*bu, ushbu*	
	Shu	

Pronouns representing anaphora in the Uzbek language are listed in the following Table 2:

1.2 Identification of Proper Nouns (NER)

Names given to certain persons or things are called proper nouns and are written with capital letters. They include:

- Names and nicknames of people: Dilmurod, Ravshan Mahmudov, Zarifa Hakimjon qizi, Cho ‘lpon, Julqunboy, etc.
- Names given to pets: Olapar, To ‘rtko ‘z, Mosh, Boychibor, Yo ‘lbars, etc.
- Geographical and astronomical names: O ‘sh, Chirchiq, Oqtosh, Kavkaz, Dnepr, Mars, Oy, Quyosh, etc.
- Names of republics and higher organizations: *O ‘zbekiston Respublikasi, Vazirlar Mahkamasi, Yevropa Taraqqiyot va Tiklanish Banki, Oliy Kengash,* etc.
- The first part of the name of scientific institutions, higher educational institutions, ministries and enterprises: *Til va adabiyot instituti, Moliya vazirligi, Bank-moliya akademiyasi*, etc.

- Names of high positions, high honorary titles: *O ‘zbekiston Prezidenti, Bosh Vazir, O ‘zbekiston Qahramon;*
- Names of factories, various organizations, collective farms, cinema, theater, books, newspapers, magazines are written with capital letters in quotation marks: *Qizil tong” fabrikasi, “Sharqyulduzi” jurnali, “Umid” jamg ‘armasi, “Muqimiy” teatri.*
- The first part of the names of historical events, dates, scientific conferences, documents: *Mustaqillik kuni, Toshkent deklaratsiyasi, Qurbon hayiti kabi.*
- The first letters of abbreviated initials: *JIDU Jahon iqtisodi va diplomatiya universiteti, BMA Bank-moliya akademiyasi*, etc.

Using the NER model, nouns are identified and categorized from unstructured text [10]. For example, personal names, organization names, place names, product names. The process of identifying NER objects is carried out in the following stages:

1. Extract information—the first step in determining NER is to extract the objects indicated in the sentence, paragraph, text. At this stage, the whole text is marked and the text border is defined. Here, the text is divided into sentences according to the capital letter system.
2. Tokenization process—chunked sentences are now tokenized within themselves.
3. Determining the limit of tokens according to the “IOB” or “BILUO” scheme and “collecting” them again—in this case, the tokens of several NERs are “united” based on the model.
4. Object search—the next process in NER is searching for NER objects in tokens.
5. Assign the correct category to identified NER objects.

Explaining the above process differently, the identified nouns are analyzed not only morphologically, but also semantically. The nouns separated from the text are defined according to their “proper noun” characteristics. Prominent nouns are analyzed according to “NER features” (occurrence in the text in capital letters, non-dictionary lexicon, addition of suffixes, object renaming, etc.). The dictionary also serves as a tool for identifying NER objects. But this base cannot be a perfectly effective solution.

NER is an NLP method that can extract the main objects in the text and divide them into predefined categories [11]. The process of identifying named objects from the text that do not exist in the dictionary, such as personal names, location names, company names, etc., is an important step in solving many NLP tasks. Named object recognition in NLP is also commonly referred to as object identification, object extraction, or object segmentation. NER object detection algorithms are the following models [12] (Fig. 1).

- analysis based on rules;
- dictionary search;
- POS tagging (morphological tagging);
- Parsing (syntactic tagging).

Fig. 1 Identification of NER objects

To understand the process of identifying NER objects from text content, consider the following sentence:

The blue ones here are nouns. Some of these nouns represent real objects that exist in the world. For example, from the above, the following nouns represent existing locations on the map: *"Toshkent"; "O 'zbekistonning"; "Markaziy Osiyodagi"*.

If we can find nouns in a text, specifically named nouns, with this kind of accuracy, we can use this information to identify a list of named objects automatically in the text in NLP. The goal of NER is to identify and label these nouns with relevant real-world concepts. (For example: Unicef is an organization (ORG); Alisher is a person's name (PER)).

NER systems do more than just look up a simple dictionary. Perhaps they use a statistical model to determine how a word appears in a sentence in the text and what type of noun that word represents [12]. The problem of identifying NER objects in the Uzbek language was analyzed by Elov and Samatboyeva [10].

1.3 Identification of Noun Phrases

In the Uzbek language, the word combinations that are freely connected in the subordinate-dominant relationship, and the main word is represented by a noun, are noun combinations. For example: *shamoldan tez, qiziqarli kitob, aqliy bilish, tezkor xotira.* It can be seen that in Uzbek, noun compounds can be expressed by words such as adjectives, numbers, adverbs, gerund or infinitive. Since solving the coreference problem is focused on specific nouns, we will focus only on compound nouns that express the modification signs of these nouns. That is, we determine the attributive relational combinations expressed by the name and the words used to describe it.

For example: *yosh, iste 'dodli, mas 'uliyatli xodim.*

1.4 Clustering

After performing the tasks in the first stage, we try to identify candidates/notes that refer to an object. We perform the process of clustering the referent candidates/notes expressed in the content into clusters corresponding to the objects indicated in the text.

For example: *Nilufarning aytishicha,* ***Azimjon****juda iltifotli yigit, chunki* ***u*** *har doim* ***Azizani*** *ishga* ***o 'zi****olib keladi, mashinaning eshigini ham* ***o 'zi*** *ochib,* ***ayolini***

ishxonasiga kuzatib qo ‘yadi. (Nilufar says that Azimjon is a very kind guy, because he always brings Aziza to work, opens the door of the car himself and escorts her to the office).

In this example, we have marked the candidate/note group that caused the coreference in separate colors. In this sentence, *Azimjon* and ***u, o‘zi****; Aziza* and ***ayol*** belong to the same object, and represent a cluster (group) among the candidates/notes representing each object. Names of nouns or objects that do not have a reference in the content are called singletons (a cluster containing only one unit).

1.4.1 Basic Terms

We will explain the terms used in the corpus in solving the issue of coreference in the example of the following sentences:

Qizaloq *bog ‘dan chiqishni istamadi, chunki* ***u*** *arg ‘imchoqlarning barchasida uchmoqchi edi. (***The girl** did not want to leave the park, because **she** wanted to swing on all flip flops).

- **Antecedent**—the object that participated in the front part of the sentence/text/content—the unit that determines the meaning of the next pronoun in the text. For example, in the example above, the meaning of the pronoun [u] is determined by [qizaloq]. Therefore, [qizaloq] is the antecedent of [u].
- **Anaphora** is a unit that refers to a person or thing (object) in the context, to what was said before (the expressed object). For example.

Tort *yesang, muzlatkishga* ***bir bo ‘lak*** *bor. (If you eat a* ***cake****, there is a* ***piece*** *in the freezer).* The word [***bir bo ‘lak***] in the second part of the sentence is an anaphoric unit referring to the word [***tort***].

- **Cataphora** is a form of anaphora, which differs in that the pronoun is expressed before the noun it refers to [13]. For example,

Negadir ***u*** *musobaqada kutilgan natijani ko ‘rsata olmadi,* ***Anvar*** *odatda bunday tadbirlarga puxta tayyorlanar edi. (For some reason,* ***he*** *could not show the expected result in the competition,* ***Anvar*** *usually prepared carefully for such events).*

In this sentence, the demonstrative pronoun [u] precedes the word (object) it refers to [Anvar].

For another example, *Jasmin ingliz tili kurslariga qatnaydi, endi* ***Malika*** *ham unga qo ‘shildi, chunki* ***u*** *ham ingliz tilini o ‘rganmoqchi (Jasmin attends English classes, and now* ***Malika*** *has joined her, because* ***she*** *also wants to learn English).*

In this sentence, [u] refers to [Malika] because this word is closer to [Jasmin] in the sentence.

In a sentence, the subject position is referred to more often than the object position. For example, ***Jasmin*** *ingliz tili kurslariga qatnaydi, endi Malika ham* ***unga*** *qo ‘shildi, chunki u ham ingliz tilini o ‘rganmoqchi.*

In this case, [her[reference refers to [Jasmine] because [Malika] is in object position and [Jasmine] is subject position.

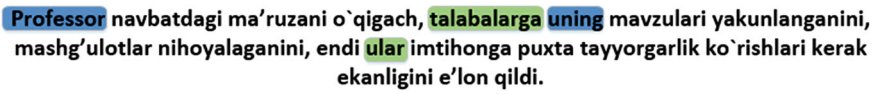

Fig. 2 The semantic connection of the words in the sentence

1.4.2 Parallelism

Nazokat Feruza *Bilan Kinoga Bordi,* ***Aziza*** *Esa* ***U*** *Bilan Savdo Markaziga Bordi (****Nazokat*** *Went to the Cinema with* ***Feruza****, and* ***Aziza*** *Went to the Mall with Her).*

In this sentence, since the word with creates a parallel structure, it refers to **Nazokat**.

It should be noted that in the following sentence, the word **"his**" refers to the word "**professor**" and not to the word "**students**" (Fig. 2).

1.4.3 Verb Semantics

Compare the following two sentences:

– *Qizim qunt bilan* ***tort*** *pishirdi va endi* ***u*** *san 'at darajasiga yetdi.*

(My daughter baked a cake diligently, and now it has reached the level of art).

Qizim *Qunt Bilan Tort Pishirdi, Endi* ***U*** *Juda Charchadi.*

(My Daughter Baked a Cake Diligently, Now She is Very Tired).

According to the structure, the above sentences have the same construction. But in each of these sentences, it is necessary to correctly determine what (which object) the pronoun "u" means.

Semantics can also tell us about the connections between the notes in a sentence.

The animal didn't cross the street because it was too tired.

The animal didn't cross the street because it was too wide.

In the first sentence, we know that the word **it** refers to an animal, and in the second sentence, it refers to the street through the words **tired** and **wide**.

1.4.4 Additional Features

There are a few more features (PNG constraints) that help create manageable classifiers:

– person (1st person, 2nd person, 3rd person);
– number (one or more);
– gender (male or female).

1.4.5 Co-reference in Uzbek Language Texts

In the traditional method, we first need to identify all the pronouns through the pipeline process. For this purpose, it is necessary to implement the process of POS tagging in the text using the Uzbek language morphoanalyzer developed by Elov, Alayev and Hamroyeva. In the second step, using the NER object detection system developed by B. Elov and M. Samatboyeva, the list of necessary NERs in the text is determined. In the next step, a syntactic analyzer and systems for identifying references to nouns and clustering coreferences are used. As a result, a set of words and phrases related to coreference in Uzbek textbooks is determined through a sequence of 5 steps.

2 Conclusion

This article describes the main conditions and stages of solving the coreference problem in Uzbek language texts. Products such as electronic libraries and language corpora that reflect the national language, display it electronically, and store it in digital format are necessary tools of today's globalization era. Processing natural language, bringing it to the level of machine language, significantly increases the viability and status of the language. In order to do this, it is necessary to develop NLP applications such as semantic analyzers, question–answer systems, and machine translation that automatically analyze Uzbek language texts. We have focused on the automatic solution of the coreference problem, which is one of the urgent issues that will be necessary in the implementation of such tasks. In our further studies, we will discuss detailed information about algorithms and models of automatic solution of this problem.

References

1. McShane M, Nirenburg S (2021) Basic coreference resolution. In: Linguistics for the age of AI. https://doi.org/10.7551/mitpress/13618.003.0007
2. Toshniwal S, Xia P, Wiseman S, Livescu K, Gimpel K (2021) On generalization in coreference resolution. In: 4th workshop on computational models of reference, Anaphora and Coreference, CRAC 2021—Proceedings of the workshop. https://doi.org/10.18653/v1/2021.crac-1.12
3. Urbizu G, Soraluze A, Arregi O (2020) Sequence to sequence coreference resolution. In: Proceedings of the third workshop on computational models of reference, Anaphora and coreference
4. Sukthanker R, Poria S, Cambria E, Thirunavukarasu R (2020) Anaphora and coreference resolution: a review. Inf Fusion 59:139–162. https://doi.org/10.1016/j.inffus.2020.01.010
5. Stylianou N, Vlahavas I (2021) A neural entity coreference resolution review. In Expert Syst Appl 168. https://doi.org/10.1016/j.eswa.2020.114466
6. Lata K, Singh P, Dutta K (2022) Mention detection in coreference resolution: survey. Appl Intell 52(9). https://doi.org/10.1007/s10489-021-02878-2

7. Cruz AF, Rocha G, Cardoso HL (2020) Coreference resolution: toward end-to-end and cross-lingual systems. Information (Switzerland) 11(2). https://doi.org/10.3390/info11020074
8. Ji D, Gao J, Fei H, Teng C, Ren Y (2020) A deep neural network model for speakers coreference resolution in legal texts. Inf Process Manag 57(6). https://doi.org/10.1016/j.ipm.2020.102365
9. Сайфуллаева Р., Менглиев Б., Қурбонова М., Боқиева Г., Юнусова З., Абузалова М. Ҳозирги ўзбек адабий тили. Ўқув қўлланмаси. Тошкент, 2005. 388 б.
10. Elov B, Samatboyeva M (2023) Identifying NER (Named Entity Recognition) objects in Uzbek language texts. Sci Innov Int Sci J 2(4)UIF-2022:8.2 | ISSN:2181–3337
11. Dobrovolskii V (2021) Word-Level coreference resolution. EMNLP 2021—2021 conference on empirical methods in natural language processing, Proceedings. https://doi.org/10.18653/v1/2021.emnlp-main.605
12. https://www.analyticsvidhya.com/blog/2021/06/part-1-step-by-step-guide-to-master-natural-language-processing-nlp-in-python/ (2021)
13. Kantor B, Globerson A (2019) Coreference resolution with entity equalization. In: ACL 2019—57th Annual meeting of the association for computational linguistics, Proceedings of the conference. https://doi.org/10.18653/v1/p19-1066