УДК 81'322

РАЗРЕШЕНИЕ АНАФОРЫ В ВЫЧИСЛИТЕЛЬНОЙ ЛИНГВИСТИКЕ

Г.Р. Еремеева, О.К. Мельникова, А.Р. Ситтикова
guzeliksanova@rambler.ru, melnikova.olga@mail.ru, sitti.alina@mail.ru

Казанский (Приволжский) федеральный университет, г. Казань, Россия

Аннотация. Статья посвящена проблеме разрешения анафоры в вычислительной лингвистике. Разрешение анафоры требуется в большинстве задач в области обработки естественного языка. Целью исследования является изучение того, как инструмент разрешения анафоры справляется со своей задачей. Проведенный анализ позволяет утверждать, что инструменты для этой задачи должны быть улучшены, а также необходимо провести больше практической работы в этой области. Результаты исследования могут найти применение в дальнейших исследованиях и попытках повысить эффективность инструментов разрешения анафоры.

Ключевые слова: лингвистика, анафора, кореферентность, разрешение анафоры, Coreference Resolver, местоимение.

Для цитирования: Еремеева Г.Р., Мельникова О.К., Ситтикова Ф.Р. Разрешение анафоры в вычислительной лингвистике // Казанский лингвистический журнал. 2018, том 1, № 1 (1). С. 57–66.

ANAPHORA RESOLUTION IN COMPUTATIONAL LINGUISTICS

G.R. Eremeeva, O.K. Melnikova, A.R. Sittikova
guzeliksanova@rambler.ru, melnikova.olga@mail.ru, sitti.alina@mail.ru

Kazan Federal University, Kazan, Russia

Abstract. The paper is devoted to the problem of anaphora resolution in computational linguistics. Anaphora resolution is required in most tasks in Natural Language Processing. The aim of the research is to investigate how the tool for anaphora resolution cope with its task. The analysis suggests that tools for the task need to be improved and more practical work in the field should be done. The results of the research can find application in further studies and attempts to improve the effectiveness of anaphora resolution tools.

Keywords: linguistics, anaphora, coreference, anaphora resolution, Coreference Resolver, pronoun.

For citation: Eremeeva G.R., Melnikova O.K., Sittikova A.R. Anaphora resolution in computational linguistics // Kazan linguistic journal. 2018, Vol. 1, No. 1 (1). Pp. 57–66.

Currently, resolution of anaphoric reference is one of the most challenging tasks in Natural Language Processing (NLP). Anaphora resolution is required in most tasks such as question-answering, information extraction, text summarization, dialogue interpretation systems, etc. [2], [8]. Thereby, to a large extent, systems require a successful anaphora resolution module. There is a wide variety of work in the area, based on various theoretical approaches – traditional, alternative or knowledge-poor. Despite the ongoing work on developing of new models and algorithmic techniques, most solutions are only approximate. Therefore, the purpose of this paper is to investigate on a newspaper article how a tool for anaphora resolution copes with this task and which types of anaphora it identifies better or worse.

In this article the Demo version of Coreference Resolver (presented at EMNLP 2008) is used. The Demo version was chosen for the lack of technical capabilities to install the full version. It includes an anaphoricity classifier trained using machine learning techniques and other features. This is a key ingredient in the performance of the used coreference classifier. After siting anaphoric references with this tool we analyzed the results by comparing them with the anaphora analysis done by human and drew the conclusion.

The etymology of the term "anaphora" goes back to Ancient Greek with “anaphora” (αναφορα) being a compound word consisting of the separate words ανα − back, upstream, back in an upward direction and φορα – the act of carrying and denoted the act of carrying back upstream.

Various definitions of anaphora have been put forward, but the classical definition is: anaphora is a cohesion (presupposition) which points back to some previous item [10].

The "pointing back" (reference) is called an anaphor and the entity to which it refers is its antecedent. The process of determining the antecedent of an anaphor is called anaphora resolution. Usually, both the antecedent and the anaphor are used as referring expressions and having the same referent in the real world, they are termed coreferential [4].

Example:

The Queen is supposed to be in England now but she was seen in France.

In this example, the antecedent is "the Queen" and the anaphor is the pronoun "she". Pay attention that the antecedent is not the noun "Queen" but the noun phrase "the Queen".

There may be cases when the anaphor and more than one of the preceding (or following) entities (usually noun phrases) have the same referent and are therefore pairwise coreferential, thus forming a coreferential chain. In such a case, we regard each of the preceding entities, which are coreferential with the anaphor(s) as a legitimate antecedent. Therefore, in such cases the task of anaphora resolution is considered successful, if any of the preceding entities in the coreferential chain is identified as an antecedent.

There are various types of anaphora [5], [1], but we shall briefly outline those that are thought to be the three most widespread types in the Computational Linguistics literature.

• Pronominal anaphora

The most widespread type of anaphora is the pronominal anaphora, which is realised by anaphoric pronouns.

Example:

Computational Linguists from many different countries attended the tutorial. They took extensive notes.

It should be pointed out that not all pronouns in English are anaphoric. For instance, "it" can often be non-anaphoric such as in the case of the previous sentence. Other examples of non-anaphoric “it” include expressions such as "It is important", "It is necessary", "It has to be taken into account". A non-anaphoric "it" is termed pleonastic [6].

• Definite noun phrase anaphora

Typical cases of definite noun phrase anaphora is when the antecedent is referred by a definite noun phrase representing either same concept (repetition) or semantically close concepts (e.g. synonyms, superordinates).

Example:

Computational Linguists from many different countries attended the tutorial. The participants found it hard to cope with the speed of the presentation.

• One-anaphora

One-anaphora is the case when the anaphoric expression is realised by a "one" noun phrase.

Example:

If you cannot attend a tutorial in the morning, you can go for an afternoon one.

A very special case is cataphora, arising when reference to an entity mentioned subsequently in the text is made. Technically said, it is not anaphora, but the type of endophora converse to it, nevertheless, it is often handled as a very special type of anaphora. Cataphora is almost always intrasentential and is usually realized by a cataphoric pronoun, located in an embedded relative clause.

Example:

By the time she got home, Elizabeth was completely wet.

Finally, we distinguish intrasentential anaphors (referring to an antecedent which is in the same sentence as the anaphor) and intersentential anaphors (referring to an antecedent which is in a different sentence from that of the anaphor) [6].

Most of the anaphora resolution systems deal with resolution of anaphors which have noun phrases as their antecedents because identifying anaphors which have verb phrases, clauses, sentences or even paragraphs/discourse segments as antecedents, is a more complicated task [3]. Typically, all noun phrases (NPs) preceding an anaphor are initially regarded as potential candidates for antecedents. Usually, a search scope has to be identified: most approaches look for NPs in the current and preceding sentence. However, an "ideal" anaphora resolution system should extend its scope of search: antecedents which are 17 sentences away from the anaphor have already been reported [4].

Coreference resolution is a popular and challenging Natural Language Processing (NLP) task that involves clustering denotative noun phrases in a document where two noun phrases are co-clustered if and only if they refer to the same entity [9]. L3M is a feature-based probabilistic structured prediction model, where each item can link to a previous item with a certain probability. L3M expresses the probability of an item joining a previously formed cluster as the sum of the probablities of multiple links connecting that item to the items inside that cluster [7].

According to the goal of the research, we have chosen article “10 Beehives Stolen From Farm” written by Jordan Otero Sisson from Hartford Courant newspaper, 3 April 2018 edition. The choice of the article was random, in order to better study the accuracy of the Coreference

Resolver analysis. Due to Online Demo Version of the anaphora resolution tool restrictions the text analysis was done by parts.

In the table below two texts are presented: in the first one, the search for anaphoric references was performed manually, and in the second one it was done by Coreference Resolver.

ANALYSIS OF COREFERENTIAL CONNECTIONS IN THE TEXT

Table 1

Human analysis	Coreference Resolver analysis
Police are investigating the theft of 10 beehives that were discovered missing from a Farmington farm last week. **[The hives], [property of Stonewall Apiary in the Hanover section of Sprague]**, were stolen from *[Sub Edge Farm]* in Farmington sometime between the end of January and March 29, Stonewall Apiary owner ***[Stuart Woronecki]*** wrote in a Facebook post Monday morning. *[Sub Edge Farm*] estimated the monetary loss at $2,000 in a separate post on *[its]* business Facebook page. [Rodger Phillips], [owner of Sub Edge Farm], deferred further comment to ***[Woronecki]*** when reached on Facebook Monday. ***[Woronecki]*** did not return an emailed request for comment. ***[Lt. Kelly Walsh of the Avon Police Department]*** said Monday that the theft was reported on Thursday and that there were no updates Monday afternoon. In the report, [Phillips] said [he] had "last seen the **[hives]** on [his] property at the end of January and didn't notice **[they]** were missing until [March 29] at approximately [4 p.m.]," ***[Walsh]*** said. ***[Walsh]*** said ***[she]*** wasn't aware of similar crimes in the area. "This is not a normal theft," ***[Walsh]*** said.	Police are investigating the theft of 10 beehives that were discovered missing from a Farmington farm last week. The hives, **[[property of Stonewall Apiary]]** in the Hanover section of Sprague, were stolen from Sub Edge Farm in Farmington sometime between the end of January and March 29, Stonewall Apiary owner Stuart Woronecki wrote in a **[Facebook]** post Monday morning. Sub Edge Farm estimated the monetary loss at 2,000 in a separate post on its business Facebook page. [Rodger Phillips], [[owner of Sub Edge Farm]], deferred further comment to [Woronecki] when reached on [Facebook] Monday. [Woronecki] did not return an emailed request for comment. ***[Lt. Kelly Walsh of the Avon Police Department]*** said Monday that the theft was reported on Thursday and that there were no updates Monday afternoon. In the report, [Phillips] said [he] had "last seen the hives on [his] property at the end of January and didn't notice they were missing until March 29 at approximately 4 p.m.," ***[Walsh]*** said. ***[Walsh]*** said ***[she]*** wasn't aware of similar crimes in the area. "This is not a normal theft," ***[Walsh***] said.

Продолжение Table 1

Human analysis	Coreference Resolver analysis
*[Both **[Woronecki]** and Sub Edge]* wrote in separate Facebook posts that *[they]* believe the theft was the work of someone with beekeeping experience. In ***[his]*** post to the Facebook group, "BeekeepersOfCT," ***[Woronecki]*** said that the hives were "in a state of disarray" and that rain, snow and mice had gotten into some of the "deeps," or boxes that contain frames of comb. ***[He]*** wrote that aside from needing a new foundation and frames, everything is reusable. The hives were closed off, but not in a locked area, ***[Woronecki]*** said in a reply to a comment on ***[his]*** post left by another user. ***[Woronecki]*** wrote that Stonewall Apiary does have insurance to protect against such losses, and that ***[he]*** has equipment to replace what was damaged. ***[He]*** speculated that the theft occurred during the warm weather in February and that at that time ***[the thieves]*** "could have gotten in and out of there in 20 minutes or less." "The thing that was most valuable in these hives is the bees. That is clearly what ***[they]*** were after," ***[Woronecki]*** wrote on Facebook.	Both **[Woronecki]** and Sub Edge wrote in separate **[Facebook]** posts that they believe the theft was the work of someone with beekeeping experience. In his post to the **[Facebook]** group, "BeekeepersOfCT", **[Woronecki]** said that the hives were "in a state of disarray" and that rain, snow and mice had gotten into some of the "deeps", or boxes that contain frames of comb. **[He]** wrote that aside from needing a new foundation and frames, everything is reusable. The hives were closed off, but not in a locked area, [Woronecki] said in a reply to a comment on **[his]** post left by another user. **[Woronecki]** wrote that Stonewall Apiary does have insurance to protect against such losses, and that [he] has equipment to replace what was damaged. [He] speculated that the theft occurred during the warm weather in February and that at that time the thieves "could have gotten in and out of there in 20 minutes or less". "The thing that was most valuable in these hives is the bees. That is clearly what they were after," **[Woronecki]** wrote on **[Facebook].**

Продолжение Table 1

Human analysis	Coreference Resolver analysis
[State entomologist Dr. Kirby C. Stafford III] said most Connecticut beekeepers are hobbyists and all are legally required to register with the Connecticut Agriculture Experiment Station. As of last year, there were **[1,689 beekeepers]** registered with over 8,000 beehives in Connecticut. Stonewall Apiary was last registered in 2011, [Stafford] said. [Stafford] said that beehive thefts are very uncommon in Connecticut. While ***[anecdotal reports]*** have been made in the past to the Connecticut Agriculture Experiment Station, [Stafford] said ***[those instances]*** were unconfirmed and only one or two were stolen hives. "It's not common by any means that I'm aware of," [Stafford] said. "It would take a lot of work to move 10 hives." [Stafford] said [he] could only speculate about motivations for such a crime, but suggested that winter losses are a possible reason. Statistics aren't in for this season, but **[beekeepers]** reported an approximate 61 percent loss on **[their]** hives last winter. Typically, losses are closer to or under 30 percent, [he] said. "[Stealing] would be a quick way to recover hives if you've lost some, rather than importing *[a nucleus colony]*, *[or startup colony]*," [Stafford] said.	State entomologist Dr. Kirby C. Stafford III said most **[Connecticut]** beekeepers are hobbyists and all are legally required to register with [**[the Connecticut Agriculture]Experiment Station].** As of last year, there were [1,689 beekeepers] registered with over 8,000 beehives in **[Connecticut].** Stonewall Apiary was last registered in 2011, [Stafford] said. [Stafford] said that beehive thefts are very uncommon in **[Connecticut].** While anecdotal reports have been made in the past to [**the [Connecticut Agriculture] Experiment Station],** [Stafford] said those instances were unconfirmed and only ***[one or two]*** were stolen hives. "It's not common by any means that I'm aware of," [Stafford] said. "It would take a lot of work to move 10 hives." [Stafford] said [he] could only speculate about motivations for such a crime, but suggested that winter ***[losses]*** are a possible reason. Statistics aren't in for this season, but [beekeepers] reported an approximate 61 percent loss on their hives last winter. Typically, ***[losses]*** are closer to or under 30 percent, he said. "[Stealing] would be a quick way to recover hives if you've lost some, rather than importing *[a nucleus colony]*, *[or startup colony]*," [Stafford] said.

Продолжение Table 1

Human analysis	Coreference Resolver analysis
While *[Sub Edge Farm]* estimated the stolen hives' value at $2,000, ***[Steve Dinsmore], [president of the Connecticut Beekeepers Association],*** said the loss is likely closer to $5,000. ***[Dinsmore***] said the cost of raising bees could be what motivated [the theft].	While **[Sub Edge Farm]** estimated the stolen hives value at $2,000, **[Steve Dinsmore], [president of** [the Connecticut Beekeepers Association]], said the loss is likely closer to $5,000. **[Dinsmore]** said the cost of raising [bees] could be what motivated the theft.
[Dinsmore] said [beekeepers] can use wildlife cameras to monitor hives, or a brand to mark [their] [boxes and frames] so [they] are easily identifiable if [they] turn up elsewhere. Still, ***[Dinsmore]*** said ***[he]*** is unaware of any recent beehive thefts in Connecticut.	**[Dinsmore]** said [beekeepers] can use wildlife cameras to monitor hives, or a brand to mark ***[their]*** boxes and frames so ***[they]*** are easily identifiable if ***[they]*** turn up elsewhere. Still, **[Dinsmore]** said he is unaware of any recent beehive thefts in Connecticut.
"[It] definitely raises some alarms," ***[Dinsmore]*** said. "It means the good feelings ***[we]*** have toward ***[our]*** fellow beekeepers are somewhat under question at this point."	"It definitely raises some alarms," **[Dinsmore]** said. "It means the good feelings ***[we]*** have toward ***[our]*** fellow beekeepers are somewhat under question at this point."
"Hopefully this isn't a trend. Hopefully whoever does [it] gets caught," ***[Dinsmore]*** added.	"Hopefully this isn't a trend. Hopefully whoever does it gets caught," **[Dinsmore]** added.
In a comment on its Facebook post, *[Sub Edge Farm]* said that replacement bees will be ordered this spring "so we are not giving up."	In a comment on its Facebook post, **[Sub Edge Farm]** said that replacement **[bees]** will be ordered this spring "so we are not giving up". The farm also wrote that it would not ask for financial support but encouraged *[the community]* to spread the word of the theft and continue to support *[its]* farmstand.
[The farm] also wrote that *[it]* would not ask for financial support but encouraged *[the community]* to spread the word of the theft and continue to support *[its]* farmstand.	A GoFundMe online fundraiser has been set up for the cause with a fundraising goal of $2,000.
A GoFundMe online fundraiser has been set up for the cause with a fundraising goal of $2,000.	
[Facebook user Darlene Redman] said in a Monday update on the site that money raised is "intended to replace the bees, equipment and possibly install some type of motion light or camera."	[Facebook] **user Darlene Redman]** said in a Monday update on the site that money raised is "intended to replace the bees, equipment and possibly install some type of motion light or camera."
[Redman] did not respond to a Facebook message requesting comment Monday.	**[Redman]** did not respond to a **[Facebook]** message requesting comment Monday.

The Coreference Resolver best defines the anaphoras associated with proper nouns, but does not always link appositions. Also this tool copes well with pronominal anaphoras with the pronouns "he" and "she", but almost never sees anaphoric references with the pronouns "they" and "it". The main problem is determination of intersententional anaphoras that are not proper nouns. In some sentences, Resolver linked two pronouns together, but did not find the right antecedent for them.

The results of comparison of these texts are in the table below (table 2). The analyzed text was estimated by different types of anaphora: pronominal, definite noun phrase, cathaphora, intrasententional and intersententional anaphora. No one-anaphora was found in the text, so it was not included in the table. Therefore, it can be easily seen that Coreference Resolver tool linked only 58% of all anaphoric references in this article.

RESULTS OF COMPARISON

Table 2

Type of anaphora	Total count of anaphoric references	Count of anaphoric references by Coreference Resolver
Pronominal	22	9
Definite noun phrase	27	21
Cathaphora	1	0
Intrasentential	24	11
Intersentential	27	19

Therefore, it can be easily seen that Coreference Resolver tool linked only 58% of all anaphoric references in this article.It can be explained by difficulty of understanding the right antecedent for the anaphor. It should be noted that Coreference Resolver showed good results (77% of cases were determined correctly) in definite noun phrase anaphoric references because of the specifics of the text: it has a lot of proper nouns that are easily linked together. Only 40% of pronominal anaphoras were determined correctly, 46% of intrasententional anaphoras and 70% of intersententional anaphoras. There is one cathaphora in the article that has not been determined, but it may be the case that was not provided for determination in the program.

The research showed that Coreference Resolver tool does not show excellent results in anaphora resolution and needs to be improved. Since choosing an appropriate tool to perform this task had to rely on availability to use only on laptop, we can assume that the full version of the program identifies anaphoric references better, but it was not possible to verify this. In this article, only one tool was used for anaphora resolution, again because of the technical impossibility of using others. This shows that further development of the practical part in anaphora resolution is necessary.

References

1. Basaldella M., Chiaradia G.& Tasso C. Evaluating anaphora and coreference resolution to improve automatic keyphrase extraction / Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers Osaka, Japan, 2016. Pp. 804–814. URL: http://www.aclweb.org/anthology/C16-1077 (accessed in April 21, 2018).

2. Batura T.V., Sokolov A.V. Razrabotka sistemy razresheniya anafory na osnove metodov mashinnogo obucheniya // Programmnye produkty i sistemy, 2017. Vol. 30. № 3. Pp. 461–468.

3. Boyarskij K.K., Kanevskij E.A., Stepukova A.V. Vyyavlenie anaforicheskih otnoshenij pri avtomaticheskom analize teksta // Nauchno-tekhnicheskij vestnik informacionnyh tekhnologij, mekhaniki i optiki, 2013. № 5 (87). Pp. 108–112.

4. Mitkov Ruslan. Anaphora resolution: the state of art. – URL: http://www.wlv.ac.uk/~le1825/ (accessed in April 24, 2018)

5. Hirst Graeme. Discourse-Oriented Anaphora Resolution in Natural Language Understanding: A Review. – URL: http://www.aclweb.org/anthology/J81-2001 (accessed in April 25, 2018)

6. Lappin Shalom, Herbert Leass. An algorithm for pronominal anaphora resolution. URL: http://www.aclweb.org/anthology/J94-4002 (accessed in April 25, 2018)

7. Samdani Rajhans. A Discriminative Latent Variable Model for Online Clustering. URL: https://cogcomp.org/page/publication_view/740 (accessed in May 07, 2018)

8. Sayed Imran Q. Issues in Anaphora Resolution. URL: https://nlp.stanford.edu/courses/cs224n/2003/fp/iqsayed/project_report.pdf (accessed in April 23, 2018)

9. Sukthanker R., Poria S., Cambria E. & Thirunavukarasu R. Anaphora and Coreference Resolution: A Review. URL: http://sentic.net/anaphora-and-coreference-resolution.pdf (accessed in May 05, 2018)

10. Yacko V.A. Algoritmy i programmy avtomaticheskoj obrabotki teksta // Vestnik Irkutskogo gosudarstvennogo lingvisticheskogo universiteta, 2012. Pp. 150–161.

Авторы публикации

Еремеева Гузель Ринатовна *– кандидат педагогических наук, доцент кафедры иностранных языков для физико-математического направления и информационных технологий Института международных отношений Казанского (Приволжского) федерального университета, г. Казань, ул. Межлаука, д.3.*
Email: guzeliksanova@rambler.ru

Мельникова Ольга Константиновна *– старший преподаватель кафедры иностранных языков для физико-математического направления и информационных технологий Института международных отношений Казанского (Приволжского) федерального университета, г. Казань, ул. Межлаука, д.3.*
Email: melnikova.olga@mail.ru

Authors of the publication

Eremeeva Guzel Rinatovna *– Candidate of pedagogical sciences, associate professor at the Department of foreign languages for Physics, Mathematics and Information Technologies, Institute of International Relations, Kazan (Volga region) Federal University, 3 Mezhlauk str, Kazan.*
Email: guzeliksanova@rambler.ru

Melnikova Olga Konstantinovna *– senior lecturer at the Department of foreign languages for Physics, Mathematics and Information Technologies, Institute of International Relations, Kazan (Volga region) Federal University, 3 Mezhlauk str, Kazan.*
Email: melnikova.olga@mail.ru

Авторы публикации

Ситтикова Алина Рафисовна *–*
студент Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета,
г. Казань, ул. Кремлевская, д. 18.
Email: sitti.alina@mail.ru

Authors of the publication

Sittikova Alina Rafisovna *– student of the Institute of Computational Mathematics and Information Technologies, Institute of International Relations, Kazan (Volga region) Federal University, 18 Kremlyovskaya str, Kazan.*
Email: sitti.alina@mail.ru

Поступила в редакцию 15.08.2018.
Принята к публикации 01.09.2019.

УДК 81’33

ЯЗЫКОВОЕ И СЕМАНТИЧЕСКОЕ МНОГООБРАЗИЕ ЭРГОНИМОВ Г. БАРСЕЛОНЫ

А.Р. Исмагилова
ismalisha@mail.ru

Казанский (Приволжский) федеральный университет, г. Казань, Россия

Аннотация. Эргонимы – важный пласт городской топонимии, единицы языка, наиболее живо и ярко отображающие изменения в лексике. Статья предлагает обзор работ, посвящённых проблеме изученности эргонимов, их классификаций, а также анализ языков-доминантов формирующих эргонимикон г. Барселоны.

Ключевые слова: лингвистический ландшафт, топоним, эргоним, глобализация, эргонимикон.

Для цитирования: Исмагилова А.Р. Языковое и семантическое многообразие эргонимов // Казанский лингвистический журнал. 2018, том 1, № 1 (1). С. 66–75.

LINGUISTIC AND SEMANTIC DIVERSITY OF THE ERGONYMS OF BARCELONA

A.R. Ismagilova
ismalisha@mail.ru
Kazan Federal University, Kazan, Russia

Abstract. Ergonyms are important layer of the city toponymy, units of language that vividly and clearly show changes in lexis. This article provides an overview of studies dedicated to the state of affairs in the research of ergonyms, their classifications, and functions and analysis of language diversity that forms the ergonimy of Barcelona.