First Steps in Automatic Anaphora Resolution in Lithuanian Language Based on Morphological Annotations and Named Entity Recognition

Voldemaras Žitkus and Lina Nemuraitė(✉)

Department of Information Systems,
Kaunas University of Technology, Kaunas, Lithuania
{voldemaras.zitkus,lina.nemuraite}@ktu.lt

Abstract. Anaphora resolution is an important part of natural language processing used in machine translation, semantic search and various other information retrieval and understanding systems. Anaphora resolution algorithms usually require linguistic pre-processing tools and various expensive resources for automatically identifying anaphoric expressions. Many smaller languages, like Lithuanian, lack such resources and tools. In this paper, an algorithm is proposed that requires only morphological annotations and recognized named entities. The paper presents experimental results showing the relevance of the solution for specific domains, and considers the further immediate ways towards dealing with the overall anaphora resolution problem for Lithuanian language.

Keywords: Anaphora resolution · Natural language processing · Named entity recognition · NER · Lithuanian language · References

1 Introduction

In Natural Language Processing (NLP), the anaphora is an expression the interpretation of which depends upon another word or phrase present in context (its antecedent or postcedent [1]. Anaphora resolution is important in semantic annotations for corpora and, consequently, in various systems, like semantic search, that use semantic annotations.

For example, "Tom skipped the school today. He was sick." Here words "Tom" and "He" form an anaphora, where "Tom" is an antecedent and "He" is an anaphoric object. Without anaphoric relationships, we would not be able to determine, why Tom skipped the school nor who was sick. In such cases, we would lose semantic information, amount of which mostly depends on the type of the text – for example, technical manuals less often tend to use anaphoric expressions than newspaper articles.

The relation between anaphoric object and its antecedent is intransitive, ireflexive and asymmetric. The interpretation of an anaphoric object requires another object (antecedent) that it refers to [2]. The order of an anaphoric object and the word that the anaphoric object refers to is important. If the anaphoric object follows the word that it refers to then that word is called "antecedent". If the word follows the anaphoric object

G. Dregvaite and R. Damasevicius (Eds.): ICIST 2015, CCIS 538, pp. 480–490, 2015.
DOI: 10.1007/978-3-319-24770-0_41

then that word is called “postcedent”, and such type of reference is called “cataphora”. Due to their similarity, and anaphora being more widely used, the distinction is usually not made. In this paper, we also do not distinguish between anaphoric and cataphoric expressions since the proposed algorithm applies to both types of references.

Anaphora resolution approaches fall into two broad categories: knowledge-rich and knowledge-poor ones [3]. While both approaches have different focus they require expensive resources like syntactic annotations and semantic information, or pre-annotated (often by hand) corpora. Smaller languages like Lithuanian language lack such resources. Therefore, an alternative that depends on existing Lithuanian language processing tools as morphological annotations and Named Entity Recognition (NER) is useful. It allows producing results while other, more expensive resources are still being created, and serves as a starting point from which anaphora resolution algorithms for Lithuanian language can progress further.

The rest of the paper is structured as follows. Sections 2 and 3 overview the application context, for which the anaphora resolution algorithm was developed, and related works. Section 4 explains a taxonomy of Lithuanian anaphoric expressions. Sections 5 and 6 present the anaphora resolution algorithm and its experimental evaluation results. Section 7 draws conclusions and presents future works.

2 Semantic Search Framework for Lithuanian Internet Corpus

The needs for automatic anaphora resolution in Lithuanian language had arisen in relation with creation of Semantic Analysis and Search Framework [4] for Lithuanian Internet corpus extracted from public portals (Fig. 1). Our semantic search framework is oriented towards answering questions, presented in Structured Lithuanian (SL) language. The framework transforms these questions into SPARQL queries and executes them in ontology populated by individuals discovered by semantic annotation tool from Internet corpus.

The Structured Lithuanian language is based on Semantics of Business Vocabulary and Business Rules (SBVR) [5]; this language was created as the result of the continuing research and currently is under further development [6–10]. SBVR SL allows specifying concepts, propositions and questions for domain under consideration in the form similar to the natural language. This language is understandable for human and interpretable by computers as it is based on the formal logics of SBVR. The Semantic Search Framework is domain-specific, capable to analyse specific domains (currently, it is directed towards analysing Politics, Business and Economy, and Public Administration domains). For example, it is possible to ask “Kokie įvykiai susiję su D. Grybauskaite?” (“What events are related with D. Grybauskaite?”) during specified time intervals. For politics domain, events mean meetings, pronouncements, agreements, etc.

The semantic search by giving questions in SBVR SL is different from keywords-based search as the Semantic Search Framework uses SBVR vocabularies for describing the chosen domain, and ontologies, obtained from (or synchronized with) these vocabularies. SBVR SL questions, transformed into SPARQL, are capable

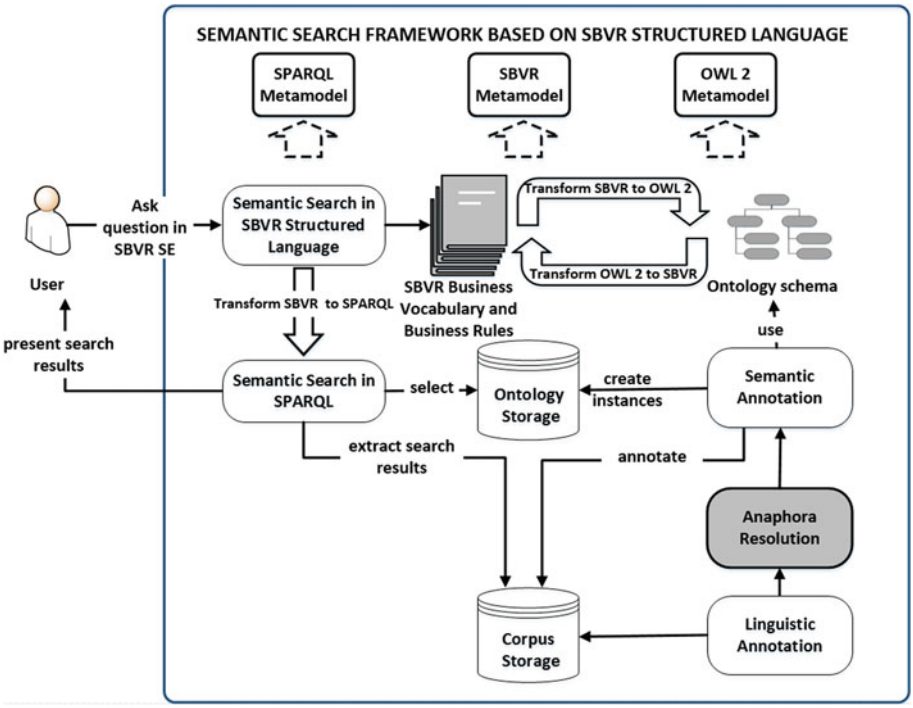

Fig. 1. Semantic search framework for Lithuanian Internet corpus [4]

to give precise results in the form of ontology individuals and their property assertions. Of course, if we want to analyse Internet contents, we have to deal with unstructured information, which must be processed by linguistic and semantic annotation tools. Semantic annotations relate recognized text fragments with individuals and their relations in ontologies; so these text fragments can be used for answering questions. The dependency on precision of linguistic and semantic annotation tools prevents from reaching full accuracy of answers; nevertheless, we have reached some encouraging results.

The anaphora resolution (Fig. 1) is just one component of this framework but it can significantly enrich the ontology population by identifying additional occurrences and links of entities, already identified after linguistic processing. To our knowledge, automated anaphora resolution tools are not available for Lithuanian language. Therefore, creation of such tools is important for further improvement of semantic search in Lithuanian language.

3 Anaphora Resolution Approaches in Other Languages

While no work has been done to solve anaphoric expressions in Lithuanian language, there are many approaches available for other languages, mostly for English. We provide the short overview of the most often cited approaches and compare their

precision (Table 1). It is important to note that the presented evaluations were performed against different corpora. Therefore, the evaluation results are not directly comparable, but they give a general understanding about the achievable results.

Table 1. Comparison of anaphora resolution approaches

Method	Foundation	Types of anaphoric expressions resolved	Precision
Hobbs	Syntactic	Main pronouns: he, she, they, it	81.8–91.7 % (depends on type of text)
BFP	Centring Theory	Pronouns (their types are not specified)	49–90 % (depends on type of text)
Left-Right Centering	Modified Centring Theory	Pronouns (their types are not specified)	72.1–81 % (depends on type of text)
RAP	Salience factors	Third person pronouns, reflexive and reciprocal anaphors	85–86 %; reaches 89 % with inclusion of statistical algorithms
Statistical approach	Probabilistic model	He, she, it and their various forms	82.9–84.2 %
Machine learning	Machine learning	Noun phrases (including pronouns)	65.5–67.3 %
UNL based approach	Universal Networking Language	Pronouns	67 %
SE-DSNL	Pattern based approach	Pronouns, but can be used for other anaphora types	81.3 %

- Hobbs algorithm is one of the earliest anaphora resolution approaches [11]. It assumes the existence of fully parsed syntactic tree with labelled nodes. The algorithm finds first pronoun that has not been analysed yet and navigates syntactic tree searching for suitable noun. When the noun is found algorithm checks if pronoun and noun agree in number and gender.
- Centring Theory based approaches (e.g., BFP [12], Left-Right Centering [13]). This theory assumes that a centre (or the focus) of the previous sentence is most likely to be pronominalized in the following sentence.
- Similar to the Centring Theory approaches, there are approaches based on salience factors (e.g., RAP [14]). Like in the Centring Theory, assumption is made that the most prominent word is likely to be an antecedent for the pronoun. Prominence is based on a number of salience factors, e.g., sentence's recency, subject's emphasis, existential emphasis, accusative emphasis, etc.
- The statistical approach [15] builds a probabilistic model that takes into account the distance between a pronoun and the candidate antecedent; the placement in the syntax tree; gender; animation; an interaction between the head constituent of the pronoun and the antecedent, and the mention count of candidate antecedents (more often mentioned antecedents are preferable).

- Machine learning approaches [16] are usually end-to-end systems that perform various NLP tasks, not only anaphora resolution. Shortcomings of other constituents of the NLP system negatively affect the anaphora resolution.
- Approach based on Universal Networking Language (UNL) [17] focuses on relationships between pronouns and their possible antecedents in previous sentences; these relationships are built on the base of semantic meaning and types of pronouns.
- SE-DSNL [18] approach attempts to determine semantic compatibility between anaphoric objects and their possible antecedents on the base of real world knowledge. For example, whether a candidate antecedent can perform the same actions as the anaphoric object.

The main difference between these methods and our proposed approach is that our algorithm is rule-based, it requires only morphological and NER annotations, and was developed for Lithuanian language.

4 Taxonomy of Anaphoric Expressions in Lithuanian Language

In our earlier work [10], we presented a taxonomy of anaphoric objects (Fig. 2) that categorizes anaphoric objects on three different levels: morphological, lexical semantics and domain semantics. The goal behind such classification was better represent actual situation where the same anaphoric expression may include the anaphoric object

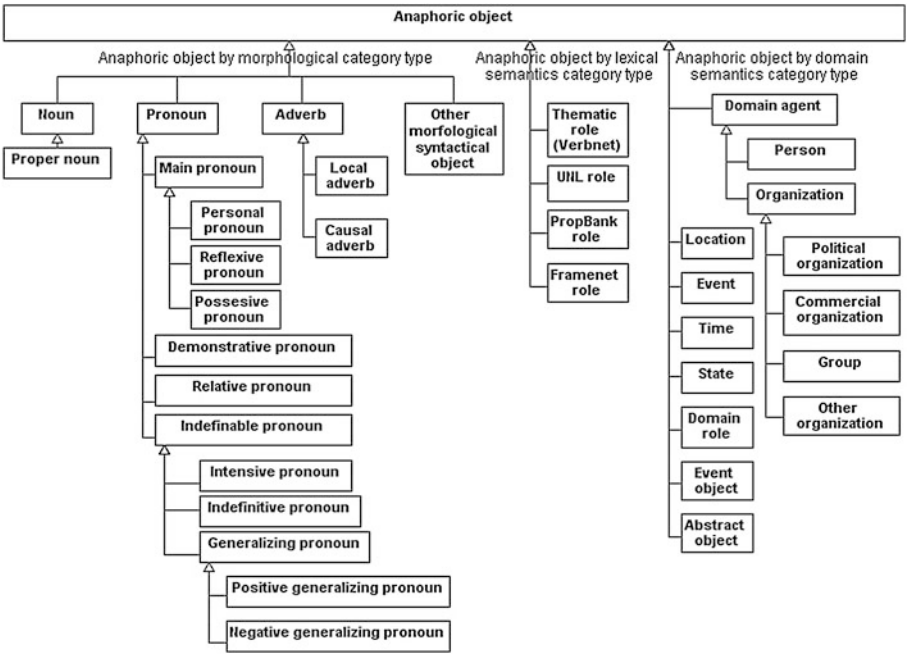

Fig. 2. Taxonomy of anaphoric expressions (adapted from [10])

that would be classified as a pronoun (morphological type), agent (lexical semantics type) and person (domain semantics type).

Some part of anaphoric relations may be detected using morphological annotations; additional relations can be found from results of lexical semantics analysis, and yet another part can be discovered from the domain semantics represented in ontology.

The generic domain semantics categories, characteristic for various domains, are adapted from [8] by extending them with state, domain role and abstract object, which are important for anaphora resolution. The "abstract object" represents such words or phrases as "person", "enterprise", "young man", etc., that can have anaphoric references. Similarly, domain roles as "president", "teacher", "politician", etc., can help discovering anaphoric relations. The morphological classification is language specific, but the lexical semantics based classification and domain semantics based classification are appropriate for other languages too.

5 Anaphora Resolution Algorithm Based on Morphological Annotations and Named Entity Recognition

In this section, we detail our proposed anaphora resolution algorithm, which was created for our Semantic Search Framework for Lithuanian Language. The algorithm was designed to provide annotations in such a way that other parts of the system can interpret its results. The algorithm was investigated on a corpus that collects articles from various Lithuanian Internet news sites focusing on political and economic matters.

Our proposed resolution method (Fig. 3) focuses on the cases where anaphoric objects are personal pronouns (subtypes of main pronouns who in turn are subtypes of pronouns in morphological categorization) and used to express persons (subtypes of domain agents in domain semantics categorization). This classification is based on our created taxonomy of anaphoric expressions as presented in Fig. 2.

The algorithm consists of the following steps:

1. Algorithm searches for the next pronoun for which anaphora resolution was not performed yet. If no new pronouns are found then we move to the ninth step.
2. Once a pronoun is found algorithm checks it against the pre-set list of invalid pronouns that usually are either pleonastic or tend not to refer to persons.
3. If the pronoun is valid, we go backwards from its position until we find a noun that is recognized as a person by Named Entity Recognition; or we reach a boundary of the sentence. If the pronoun is invalid, we return to the first step.
4. If a suitable noun is not found in the current sentence, we move backwards to the next sentence and perform the same search. This cycle continues until we either find a suitable noun, or until we pass X sentences backwards from the pronoun.
5. If we reach limit X then we move Y sentences forward from the pronoun searching for a suitable noun.
6. If we pass Y sentences forward without finding a suitable noun then the algorithm cannot determine a suitable antecedent and we return to the first step.

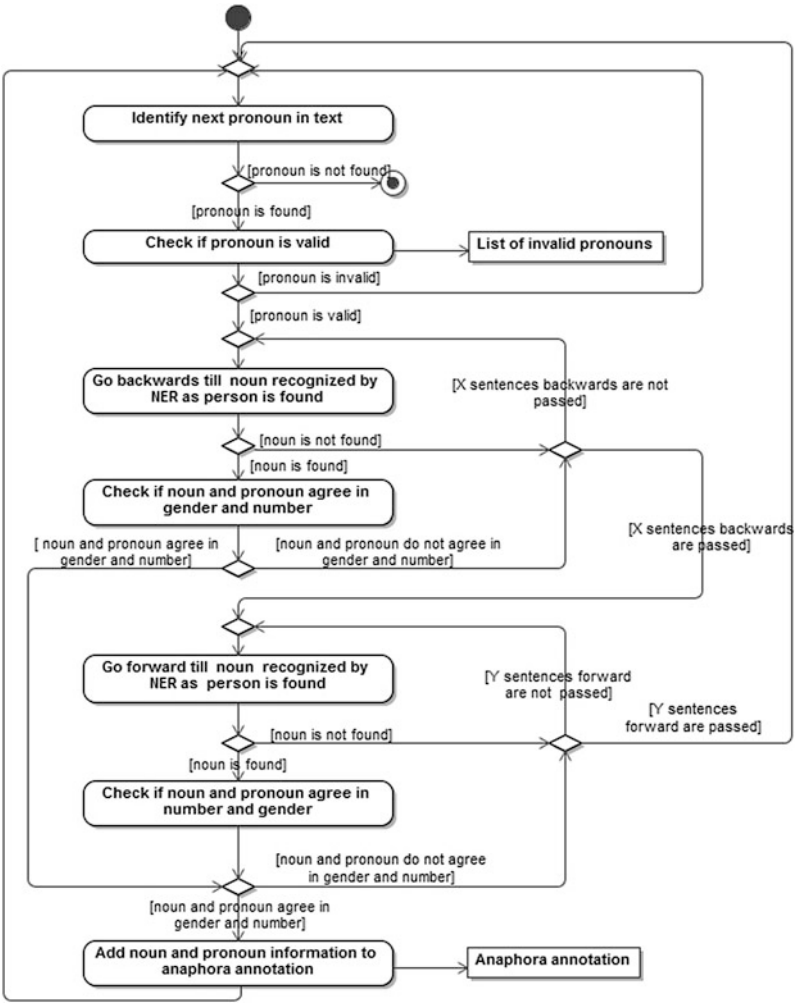

Fig. 3. Anaphora resolution method based on morphological and NER annotations

7. If during the fourth or fifth step we find a suitable noun then we determine if it agrees in number and gender with the pronoun. If it does not agree then we return to the fourth or fifth step.
8. If noun and pronoun agree in number and gender then their pair is added to anaphora annotations and we return to the first step.
9. If there is no remaining pronouns in the text, the algorithm finishes its work.

As can be seen from the detailed steps, the algorithm can be considered naive since it takes the first suitable noun that agrees in a number and gender as an antecedent, and the alternatives are not considered.

The list of invalid pronouns includes the following pronouns: kitas, tas, koks, visas, kuris, šis, joks. It has been observed that our algorithm usually is unable to find any suitable candidates for these pronouns; therefore, they are skipped in order to increase processing speed, which is relevant when working with large corpora. In addition, the omission of invalid pronouns makes the minimal impact on recall and precision. The algorithm can resolve the remaining pronouns, though it also depends on the domain under consideration. In other domains, the uses of pronouns and nouns representing persons might differ.

In our experiments, we have determined that three sentences backwards (i.e., X = 3) and one sentence forward (Y = 1) have produced the best results. These evaluations might wary for different languages and different types of texts. In total, we cover 4 sentences around anaphoric objects; the priority is given to antecedents since more steps backward is reasonable, and moving backwards is more effective.

In the following, we provide some examples on how algorithm operates on the corpus of Politics and Economy domains, annotated by morphological and NER annotation tools.

- Dalia Grybauskaitė nuvyko į Vilnių. Ji pasveikino vilniečius su šventėmis.

First, we identify pronoun "Ji". Since it is at the start of the sentence, we do not analyse remaining parts of the sentence and move one sentence backwards. At the next sentence, we start from the right and move left towards its beginning. First named entity we encounter is "Vilnių", but since it is recognized by NER as a location and not as a person we discard it, and move further to the left. Next named entity that we encounter is "Dalia Grybauskaitė", which is recognized by NER as a person. In this case, we determine the grammatical compatibility between the noun phrase (which consists of two nouns) and the pronoun. Both are singular and of female gender, therefore algorithm pairs them as the anaphoric object and its antecedent, and does not search for further candidates.

Another example:

- Kiek mažiau nei tikėtasi mokesčių mokėtojai papildė biudžetą ambicingai LR Finansų ministerijos suplanuotomis pajamomis iš akcizų (2 proc. mažiau), PVM (4 proc. mažiau), prabangaus nekilnojamojo turto mokesčio (71 proc. mažiau). Pastebėtina, kad prabangaus nekilnojamojo turto mokesčio surinkimo planas 2013 metams buvo 17 mln. litų, nepaisant to, kad 2012 m. šio mokesčio sumokėta mažiau nei 4 mln. litų (2013 m. surinkta beveik 5 mln. litų). GPM surinkimą labiausiai lėmė minimalaus mėnesinio atlyginimo (MMA) padidinimas: "Kiek man teko analizuoti, padidinus MMA tik nedidelė dalis Lietuvos įmonių sumažino etatą ar atleido darbuotojus, o tai lėmė nemažą papildomą indėlį į valstybės biudžetą" teigė Ž. Mauricas.

In this case, we first identify the pronoun "man" (the literal English translation "for me") and repeat the same steps as in the previous example. In the third sentence from the identified pronoun (the first one in the example) we find "LR finansų ministerijos" entity, which was recognized by NER as an organization and not as a person. We do not find any more entities moving backwards; therefore, we move back to our pronoun and proceed forward. The first entity we find is "Lietuvos", which is a location. We

continue moving right until we locate "Ž. Mauricas" entity, which is recognized as a person. Since the pronoun "man" is ambiguous in gender (it can refer to both female and male persons), we compare the pronoun and the noun phrase only in a number. Both are singular; therefore, we pick "Ž. Mauricas" as a postcedent of anaphoric object "man".

If any person entity would not be present in the last analysed sentence then we would not look further for a possible postcedent and pronoun "man" would be left unresolved.

Most of the named entities that are recognized by NER as persons are singular, but sometimes families are mentioned, e.g., Paulauskai, Zuokai. Due to such cases, it is important to check for agreement in number between nouns (noun phrases) and pronouns.

6 Experimental Evaluation of the Anaphora Resolution Algorithm

The purpose of the experiment was to evaluate our proposed algorithm against the corpus of Politics and Economy domains collected from Lithuanian Internet news sites in the environment of the Semantic Search framework. The evaluation was made by analysing five hundred articles that were randomly selected from around 400 thousands of collected articles.

Precision and recall are most widely used criteria for evaluation of anaphora resolution approaches. Recall R determines the percentage of anaphoric expressions F correctly resolved by the algorithm from the total number T of anaphoric expressions presented in the text (1). Precision P determines the percentage of correctly resolved anaphoric expressions C from the number of resolved anaphoric expressions F (2):

$$R = C/T \tag{1}$$

$$P = C/F \tag{2}$$

Results of the experiment are presented in Table 2.

Table 2. Results of experimental investigation of anaphora resolution algorithm performed against the subset of Politics and Economy corpus

Number of articles	T (Actual number of anaphoric expressions)	F (Number of anaphoric expressions resolved by algorithm)	P (Number of anaphoric expressions correctly resolved by algorithm)	R (Recall)	P (Precision)
500	2352	1954	1446	61 %	74 %

Considering limitations of tools and resources that we have used for pre-processing texts and implementing the algorithm, we think that results are encouraging, but the following threats to validity must be taken into account:

- Anaphoric objects that refer to named entities that NER recognizes as persons are just one small subset of possible anaphoric expressions.
- Most of the articles in the investigated corpora are taken from news portals that focus on politics and economics. The most of articles of this type could be described as collections of quotations from various politics, economists or business participants. Such texts have many named entities that can be identified as persons. In other types of texts, named entities are less often used and the algorithm would be less effective.
- Mistakes that were made due to errors in morphological or NER annotations, e.g., incorrectly identified genders of persons, were fixed by hand. Without these adjustments, the recall decreases by approximately 9 % and precision by around 4 %.
- Some articles did not have any anaphoric expressions at all or did not have pronouns referring to named entities recognized as persons. Such articles were removed from the sample set and new ones were randomly picked to replace them.
- Analysing all collected Politics and Economy corpus, we have noticed that the algorithm has not identified any anaphoric expressions in approximately 30 % of the articles. We believe that the majority of them have no pronouns referring to named entities that our algorithm could identify. The reason may be the specifics of articles, or imperfection of NER or morphological annotation tools. However, at this time we lack resources and means to validate this assumption.

7 Conclusions and Future Works

In this paper, the anaphora resolution approach was proposed for Lithuanian Internet corpus collected from news sites focusing on political and economic matters. The algorithm depends only on morphological annotations and named entity recognition, and is the only possible way towards the overall anaphora resolution problem for small languages until they have no more sophisticated linguistic pre-processing tools and resources required for this purpose.

While the algorithm provides the precision, comparable to other analysed resolution approaches, it has numerous shortcomings and limitations: it is domain specific, capable resolve just a small subset of anaphora types and was experimentally investigated for the relatively small subset of articles. The future work is directed towards investigating possibilities to adapt the similar solutions for other relevant domains and creating more sophisticated anaphora resolution algorithms using emerging tools and resources for Lithuanian language that currently are under development and will appear at the nearest future. The Semantic Search Framework for Lithuanian Internet corpora provides the favourable environment for creation and perfection of such tools, which would allow dealing with abundant information amounts in our virtual space using our native Lithuanian language.

References

1. Mitkov, R.: Anaphora Resolution. Longman, London (2002)
2. Elango, P.: Coreference resolution: a survey. Technical report, University of Wisconsin-Madison, USA (2005)
3. Mitkov, R., Lappin, S., Boguraev, B.: Introduction to the special issue on computational anaphora resolution. Comput. Linguist. **27**(4), 473–477 (2001)
4. SemantikaLT: Syntactic-semantic analysis and search system for lithuanian internet, corpus and public sector applications (2012–2014), no VP2-3.1-IVPK-12-K (2014)
5. OMG: Semantics of Business Vocabulary and Business Rules (SBVR). SBVR 1.2, version 1.2, OMG Document Number: formal/2013-11-04, pp. 1–292 (2012)
6. Sukys, A., Nemuraite, L., Sinkevicius, E., Paradauskas, B.: Querying ontologies on the base of semantics of business vocabularies and business rules. In: Information Technologies' 2011: Proceedings of the 17th International Conference on Information and Software Technologies, IT 2011, pp. 247–254, Kaunas, Lithuania, 27–29 April 2011
7. Sukys, A., Nemuraite, L., Paradauskas, B.: Representing and transforming SBVR question patterns into SPARQL. In: Skersys, T., Butleris, R., Butkiene, R. (eds.) ICIST 2012. CCIS, vol. 319, pp. 436–451. Springer, Heidelberg (2012)
8. Bernotaityte, G., Nemuraite, L., Butkiene, R., Paradauskas, B.: Developing SBVR vocabularies and business rules from OWL2 ontologies. In: Skersys, T., Butleris, R., Butkiene, R. (eds.) ICIST 2013. CCIS, vol. 403, pp. 134–145. Springer, Heidelberg (2013)
9. Karpovic, J., Krisciuniene, G., Ablonskis, L., Nemuraite, L.: The comprehensive mapping of semantics of business vocabulary and business rules (SBVR) to OWL 2 ontologies. Inf. Technol. Contr. **43**(3), 289–302 (2014)
10. Zitkus, V., Nemuraite, L.: Taxonomy of anaphoric expressions as a starting point for anaphora resolution in Lithuanian corpus. In: IVUS 2014, pp. 177–182, Lithuania (2014)
11. Hobbs, J.R.: Resolving pronoun references. In: Grosz, B., Sparck-Jones, K., Webber, B. (eds.) Reading in Natural Language Processing, vol. 99, pp. 339–352. Morgan Kaufmann Publishers Inc., San Francisco (1986)
12. Tetrault, J.R.: A corpus-based evaluation of centering and pronoun resolution. Comput. Linguis. **27**(4), 507–520 (2001)
13. Byron, D. K.: Resolving pronominal references to abstract entities. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 80–87, Philadelphia, USA (2002)
14. Lappin, S., Leass, H.J.: An algorithm for pronominal anaphora resolution. Comput. Linguis. **20**(4), 535–561 (1994)
15. Ge, N., Hale, J., Charniak, E.: A statistical approach to anaphora resolution. In: Proceedings of the Sixth Workshop of Very Large Corpora, pp. 161–170 (1998)
16. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Comput. Linguis. **27**(4), 521–544 (2001)
17. Balaji, J., Geetha, T.V., Parthasarathi, R., Karky, M.: Anaphora resolution in Tamil using universal networking language. In: Proceedings of the Indian International Conference on Artificial Intelligence, IICAI-2011, Karnataka, India (2011)
18. Fischer, W.: Linguistically motivated ontology-based information retrieval. Doctoral dissertation, University of Augsburg, GER (2013)