Automated Anaphora and Co-reference Resolution for Lithuanian Language Combining Results from Different Text Analysis Stages

Voldemaras Žitkus and Lina Nemuraitė

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

Abstract. The goal of the research is to make first steps for automated anaphora and co-reference resolution in Lithuanian language with respect to limited pre-processing tools and resources, by combining concepts and algorithms from different text analysis phases for this purpose. Existing resolution methods are created for major languages, e.g., English, and usually are language-specific. On the base of analysis of existing methods, a taxonomy of anaphoric objects is created and initial algorithms are proposed for solving anaphoras and co-references in Lithuanian language.

Keywords: anaphora resolution, co-references, natural language processing, annotation, Lithuanian language.

1 Introduction

While amounts of available information are rapidly increasing, research in Natural Language Processing (NLP) field is becoming more and more important. Unfortunately, the most of the NLP work has focused on English and other major languages leaving this field underdeveloped for smaller languages. Due to this situation, the Lithuanian language lacks mature NLP tools and resources while some parts of NLP process have not been researched at all. Anaphora and co-reference resolution is one of such cases for Lithuanian language.

In NLP, the **anaphora** is an expression interpretation of which depends on another expression in context [1]. Anaphora relation between the anaphoric object and its antecedent is an intra-linguistically determinable relation. It is nor transitive, nor reflexive, nor symmetric one [2]. The interpretation of an anaphoric object requires another object (antecedent) that it refers to, e.g.:

- Tom skipped the school today. He was sick.

The relationship between “He” and “Tom” is called an anaphora. In this case, “He” is an anaphoric object that refers to its antecedent “Tom”. Without being able to solve anaphoric expressions, we would not know why Tom skipped the school nor who was

sick. This information is very important when we try to extract semantic information from various texts.

Terms “anaphora” and “co-reference” are often used together or mistaken for each other. Sometimes, anaphoric and co-referential relations can coincide, but it is not always true. The **co-reference** is the equivalence relation between two text items (words or phrases, having the same meaning) [2]. E.g., for being referents, “lecturer” and “Mark Smith” should represent the same person. The co-reference often requires access to extra-linguistic information (the additional knowledge about the world).

The wider problem with anaphora and co-reference resolution is in the fact that even for major languages this process remains semi-automated what is entirely unacceptable to desirable analysis of the existing textual information. This is caused by imperfection of pre-processing methods and tools, needed for preparing texts for anaphora and co-reference resolution, and the lack of reliable resources, e.g., annotated corpora for resolution algorithms, based on machine learning, etc.

Research questions. This research is devoted for making the first steps in filling the gap in anaphora and co-reference resolution in Lithuanian language. It raises the following research questions:

1. Can existing anaphora and co-reference resolution methods, designed for other languages, be adapted to Lithuanian language? Can quality assessments of these algorithms be comparable with assessments of those created for major languages?
2. What automated methods and algorithms can be developed with current availability of pre-processing tools and resources in Lithuanian language?
3. How semantic information can be increased with additional (not limited to anaphora) co-reference resolution?

The research methodology is based on the Design Science Research and Information System Research Framework defined by Hevner et al. (2004) [3]. Analysis of relevant research works is being done in anaphora and co-reference resolution field for other languages. On the base of analysis made, existing methods are being adapted and new ones suitable for Lithuanian language are being created. Experiments will be performed for evaluating and improving developed methods. Resulting work will supplement the existing body of knowledge and serve as a foundation for future works on automated solving anaphoric expressions and co-references in Lithuanian and, possibly, other languages.

The rest of the paper is structured as follows. Section 2 overviews the related works. Section 3 presents the main idea of this research and provides some initial results that have been achieved. Section 4 draws conclusions and presents future works.

2 Literature review of the problem domain and related solutions

This section provides analysis of various anaphora resolution methods that were analysed in this dissertation.

Syntax based approaches. One of the earliest anaphora resolution methods was

proposed by Hobbs in 1977 [4] (often called as Hobbs's naive algorithm). Despite being the old method, it is still referenced and measured against today. The algorithm is based on fully parsed syntactic tree, finding a pronoun and navigating through the syntactic tree to determine its possible antecedent (noun). When candidate is found, the agreement in gender, number, etc., between pronoun and noun is determined on the base of morphological and real world knowledge. If the agreement is met then noun is selected as the antecedent for the pronoun, otherwise algorithm looks for another candidate. This approach encounters problems when there are several possible candidates. In such case, the algorithm would pick the first one while the other one might be correct.

Centring theory (CT). Centres link one utterance with other utterances in discourse. Each utterance has one backward-looking centre and a number of possible forward-looking centres that a particular utterance has evoked. Forward-looking centres are ranked by discourse salience and grammatical rules; the highest rated centre is called the preferred centre [5]. Brennan et al. presented one of the most known approaches (often called as BFD) that utilize CT in 1987 [6]. Tetreault proposed an alternative for this approach in 1999 (called Left-Right Centering) [7].

Salience factors. While salience plays a role in most of the approaches, usually it is not considered as the main criteria for anaphora resolution. Notable exception is RAP (Resolution of Anaphora Procedure) algorithm introduced by Lappin and Leass in 1994 [8]. Only gender, number and person of possible antecedents is taken into consideration. With each new sentence, weights of salience factors are degraded by a factor of 2. Precise weights were reached after experimentation and numerous adjustments.

Semantic information of Universal Networking Language (UNL)**. A**naphora resolution strategies based on UNL were proposed for Tamil language [9]. UNL represents semantic information of natural language texts in hyper-graphs of concepts and 46 types of relationships. Anaphoric expressions are resolved based on the types of relationships between nodes, similarly to centring and activation theories.

Semantically Enhanced Domain Specific Natural Language (SE-DSNL)**.** This approach is targeted at NLP purposes in general but can also be used for rather simplistic anaphora resolution [10]. It uses only two features (distance measuring in syntax tree and semantic compatibility) and focuses only on pronouns.

Statistical methods. One of the earliest statistical approaches was proposed by Ge et al. in 1998 [11]. The approach considers various factors for resolving anaphoric relations and investigates the relative importance of these factors while adding them incrementally.

Machine learning. First learning system to achieve comparable results with other approaches was presented by Soon et al. [12]. Their system includes tokenization and segmentation, morphological processing, part of speech tagging, noun phrase identification, Named Entity Recognition (NER), nested noun phrase extraction, and semantic class determination. In order to improve learning capabilities of the engine, authors introduced 12 feature vectors. Ng and Cardie expanded this work [13].

Comparison of approaches. The comparison of analysed resolution methods is presented in Table 1. The precision that was reported in the original research is only

given. Recall is not given since some of the methods did not provide its evaluation. The evaluations were not done against the same corpora; therefore, their results are meant to give a general idea of the state of anaphora resolution.

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%

3 Preliminary ideas of the proposed approach and the initial results

3.1 Preliminary ideas and the principal schema of the approach

The goal of the research is to make first steps for automated anaphora and co-reference resolution in Lithuanian language with respect to limited pre-processing tools and resources, by combining concepts and algorithms from different text analysis phases for this purpose. In order to reach the goal, the following tasks were stated:

1. Analyse current methods and resources used for anaphora resolution in English and other major languages;
2. Develop rules and algorithms for anaphora and co-reference resolution in Lithuanian language;
3. Implement rules and algorithms for anaphora and co-reference resolution suitable to improve semantic analysis and search in Lithuanian text corpora;
4. Conduct experiment for evaluating suitability of created rules and algorithms;
5. Evaluate developed method with recall and precision measures as main criteria.

The principal schema of the approach is presented in Fig. 1. Currently, anaphora resolution algorithms can be based on morphological annotations and entities, recognized by Named Entity Recognition (NER) algorithms, whereas existing syntactic annotation tools for Lithuanian language have not reached the sufficient quality yet.

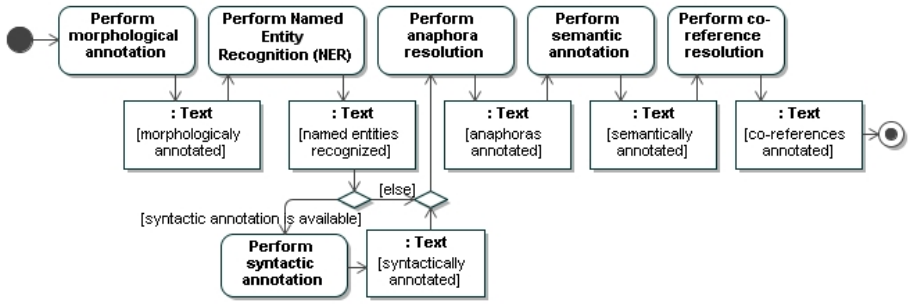

Fig. 1. The principal schema of the approach

Co-reference resolution algorithms can be applied after semantic annotation. There are more possibilities for discovering co-references, but they also are based on existence of pre-processing methods, such as, e.g., Semantic Role Labeling, so currently they are beyond the scope of this research.

3.2 Taxonomy of anaphoric expressions

This research combines multiple approaches to anaphora taxonomy by extending the main morphology-based taxonomy with additional generalization sets for providing the better coverage on the anaphora phenomenon [14]. The distinction between categories of lexical semantics and domain semantics allows identifying anaphoric expressions from multiple viewpoints.

The created taxonomy reflects the actual situation that the same anaphoric object may 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 semantic 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 extended 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 be used for discovering anaphoric relations. Morphological classification is language specific, but lexical and domain semantic classification can be used for other languages too.

3.3 Anaphora resolution algorithm based on morphological and NER annotations

This section presents the proposed anaphora resolution algorithm (Fig. 2), which was created for Semantic Search Framework for Lithuanian Language. The algorithm was investigated on a corpus that collects articles from various Lithuanian Internet news sites focusing on political and economic matters.

Proposed resolution method 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).

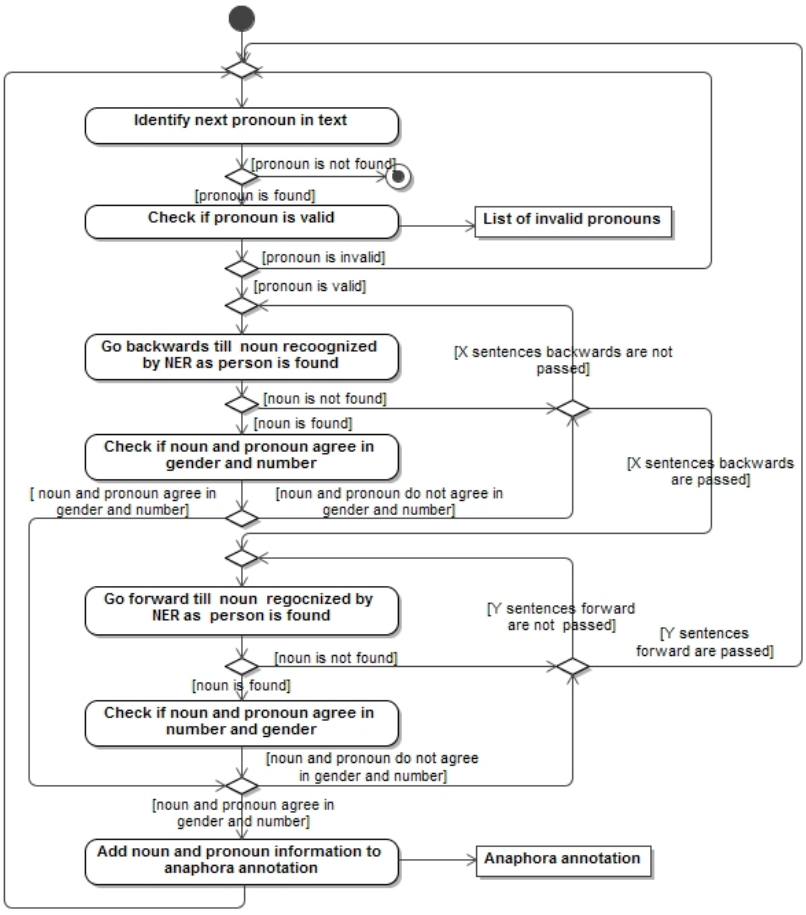

Fig. 2. Anaphora resolution method based on morphological and NER annotations

The algorithm searches for the valid pronoun for which anaphora resolution was not performed yet, and checks it against the pre-set list of invalid pronouns that usual-

ly are either pleonastic or tend not to refer to persons. If the pronoun is valid, we go backwards until we find a noun that is recognized as a person by NER. If a suitable noun is not found, we move backwards to the next sentence and perform the same search until we either find a suitable noun, or until we pass X sentences backwards from the pronoun; then we move forward Y sentences from the pronoun searching for a suitable noun.

If we find a suitable noun then we determine if it agrees in number and gender with the pronoun. If noun and pronoun agree in number and gender then their pair is added to anaphora annotations and we return to the first step.

The algorithm can be considered naive since it takes the first suitable noun that agrees in a number and gender as an antecedent (or postcedent), and the alternatives are not considered. The evaluation of the algorithm was done against corpora of 500 Internet news portal articles focusing on politics and economics. Algorithm managed to achieve 61% recall and 74% precision.

3.4 Co-reference resolution algorithm

Co-reference relation means relation between equivalent objects. In the proposed approach, equivalent objects are identified after semantic annotation (Fig. 2), during which named entities, having the same meaning but, possibly, the different representation form, are marked as different individuals.

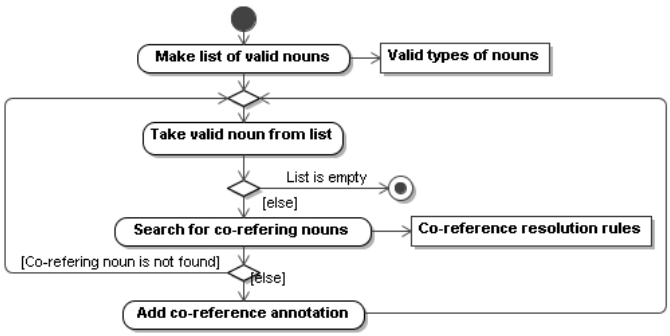

Fig. 3. Algorithm for resolving of co-references

Currently developed intertextual co-reference resolution algorithm merges such individuals into a single entity. Here, "Valid types of nouns" is a list of nouns that algorithm can resolve, e.g., persons, locations, organizations; "Resolution rules" is a list of rules that are valid for specific type (or several types) of nouns. The algorithm was tested for entities, having various modifications of their names, e.g., John Smith, J. Smith, J. S., Mr. Smith, John Smith'as, John Smithas, etc., co-refering to the same entity. The experiment was conducted with 277784 articles having 3058015 individuals. After merging, the number of individuals has decreased till 77532 (i.e., about 39 times). Unfortunately, due to the early stage of development we currently cannot provide the evaluation of precision and recall of the proposed algorithm.

4 Conclusions and future works

The paper presents the ideas and initial results after 2 years of research. The contribution of this research is the created taxonomy of anaphoric objects and algorithms for automated anaphora and co-reference resolution in Lithuanian language. Its uniqueness is in the fact that anaphoric relations and co-references are identified from multiple viewpoints via analysing categories of both lexical semantics and domain semantics. Anaphora and co-reference resolution algorithms are combined from different stages of the text pre-processing process. The research is done in the very early stage of coping with anaphora and co-reference resolution problem in Lithuanian language, with respect to imperfect pre-processing algorithms and limited resources. Therefore, the analyzed methods for other languages could not be adapted. However, the assessments of our algorithms are comparable with assessments of those created for major languages.

The future work is directed towards creating more sophisticated anaphora and co-reference resolution algorithms using emerging tools and resources for Lithuanian language that are being developed simultaneously.

References

1. Mitkov, R.: Anaphora Resolution. Longman, London (2002)
2. Van Deemter, K., Kibble, R.: On coreferring: Coreference in MUC and related annotation schemes. Computational Linguistics 26(4), pp. 629–637 (2000)
3. Hevner, A.R., March, S. T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Quarterly, 28(1), 75–105 (2004)
4. Hobbs, J.R.: Resolving Pronoun References. In: Grosz, B., Sparck-Jones, K., Webber, B. (eds.) Reading in Natural Language Processing, 99, pp. 339–352, Morgan Kaufmann Publishers Inc. (1986)
5. Kibble, R.: A Reformulation of Rule 2 of Centering Theory. Computational Linguistics, 27(4), 579–587 (2001)
6. Brennan, S.E., Friedman, M.W., Pollard, C.J.**:** A Centering Approach To Pronouns. In: Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 155–162, Philadelphia, USA (1987)
7. Tetreault, J.R.: A Corpus-Based Evaluation of Centering and Pronoun Resolution. Computational Linguistics, 27(4), 507–520 (2001)
8. Lappin, S., Leass, H.J.: An Algorithm for Pronominal Anaphora Resolution. Computational Linguistics 20(4), 535–561 (1994)
9. 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)
10. Fischer, W.: Linguistically Motivated Ontology-Based Information Retrieval. Doctoral dissertation, University of Augsburg, GER (2013)
11. 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)
12. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics 27(4), 521–544 (2001)
13. Ng, V., Cardie, C.: Improving Machine Learning Approaches to Coreference Resolution.

In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 104–111, Philadelphia, USA (July 2002)
14. Zitkus, V., Nemuraite, L.: Taxonomy of anaphoric expressions as a starting point for anaphora resolution in Lithuanian corpus. Informacinės technologijos (IVUS 2014), Kaunas, Technologija. pp. 177–182 (2014).