Using Dynamic Predicate Logic for Pronominal Anaphora Resolution in Russian Texts

Nikita Gerasimov
St. Petersburg State University
7-9, Universitetskaya nab.
St. Petersburg, Russia
n@tariel.ru

Evgeny Pyshkin
Peter the Great St. Petersburg
Polytechnic University
29 Polytechnicheskaya st.
St. Petersburg, Russia
pyshkin@icc.spbstu.ru

ABSTRACT

This study is focused on semantically-linked words detection in natural language written documents with particular attention paid to pronominal anaphora resolution in Russian language. The objective of this study is to investigate whether the approach based on using dynamic predicate logic (DPL) is appropriate to formalize a Russian pronominal anaphora and to resolve anaphoric connections and antecedents in Russian texts. As a result of the experiments that we arranged, we realized that currently the DPL model and related algorithms are not adequate to be directly applied to anaphora resolution in Russian texts due to high computational complexity and low anaphora detection accuracy.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Theory

Keywords

Pronominal anaphora, Predicate logic, PLA, DPL, NLP

1. INTRODUCTION

In linguistics, *anaphora* is the use of a language expression that can be interpreted correctly by taking into consideration its dependency on another expression (the latter being its *antecedent* or *postcedent* depending on whether it is used after or before the referred sentence). For AI systems such references are the very problem on the score of understanding anaphoras and antecedents as differenct objects. This case impairs quality of knoledge extraction and AI processing. Anaphora identification is one of the complex tasks in the domain of semantic analysis and natural language processing (NLP). The problem if particularly challenging for the current human-centric computer systems that use many different ways of human oriented interaction and require support for NLP.

Anaphora resolution could be considered as a special case of a general problem of recognition of referencing one semantic object by using different words and implicit connections. For example, learning cognitive synonyms can significantly improve the quality of web search and query expansion [12, 7]. A particularly difficult case of anaphora resolution is the case in which the reference word is a pronoun (pronominal anaphora).

Despite since about 1970s we know many approaches and algorithms developed for pronominal anaphora resolution (see the exhaustive study of Ruslan Mitkow [10], for example), the problem isn't totally fixed even for the English language [3, 6]. In contrast to English, developing algorithms for Russian language anaphora resolution seems to be still an emerging area. Among recent works we could cite the *An@phora* system [8] based on machine learning techniques and knowledge engineering formalisms as well a rule based approach described in [9].

In this short paper we report an attempt to use the Paul Dekker's approach described in [2] for pronominal anaphora resolution in Russian texts. We also report some preliminary results of testing our implementation against a corpus of Russian texts.

IWAIT '15, Oct. 8 – 10, 2015, Aizu-Wakamatsu, Japan.

2. DYNAMIC PREDICATE LOGIC

Dynamic predicate logic (DPL) is a dynamic semantic interpretation of the first-order logic language for the purposes of defining a compositional theory of discourse semantics [5].

2.1 DPL in Brief

Lets us illustrate the difference between the first-order logic and the DPL by an example. Look at the the following sentence:

Maria borrowed the textbook from her professor.

By using the first-order logic we can represent the semantic of this sentence by the following construction:

$$\exists x[B(x) \wedge \exists y(P(y,x) \wedge T(m,y,x))] \qquad (1)$$

where $B(x)$ is the textbook, $P(y,x)$ represents the fact that professor y owns the book x, and $T(m,y,x)$ represents the fact that Maria m borrowed x from y.

The anaphoric connection of *her* to *Maria* is represented correctly since all constructs in the conjunction from the equation (1) have the same scope (or, by using terms from [5] they are bound by the same existential quantifier [...]).

Let's add the following phrases:

> The textbook$_{(x)}$ was full of comments and remarks. Surely, he$_{(y)}$ was reading it$_{(x)}$ very attentively.

If we analyze the latter discourse independently of the above introduced sentence, it would be hard to resolve cross-sentential anaphoric references from *he* to *the professor*, as well as from *it* (the textbook) to that exact textbook borrowed by Maria.

Under the DPL model, the values x and y are, in a sense, "saved" in order to expand the search of suitable values for the whole domain D within the framework of the model $M = < D, I >$, D being the domain of individuals under discussion, while I – an interpreting function.

Thus, DPL is a kind of first-order logic where the following rules are introduced:

THEOREM 1. *Egli theorem*

$$(\exists x\varphi \wedge \psi) \leftrightarrow \exists x(\varphi \wedge \psi)$$

THEOREM 2. *Egli corollary*

$$(\exists x\varphi \rightarrow \psi) \leftrightarrow \forall x(\varphi \rightarrow \psi)$$

In order to see how the theorems (1–2) could be used, let's consider the following classical whistler-examples:

> A Kid is Going home. He is Whistling.
>
> A Kid who is Going home is Whistling.

By applying the above rule, both examples are being translated into the following predicate logic formula:

$$(\exists x(Kx \wedge Gx) \wedge Wx) \leftrightarrow \exists x((Kx \wedge Gx) \wedge Wx) \quad (2)$$

Moreover, unlike to classic logic, DPL offers idempotence and commutativity properties: $\phi \wedge \psi \leftrightarrow \psi \wedge \phi$ or $\phi \leftrightarrow \psi \vee \phi$ is false if the interpretation changes from ϕ to ψ. Again, if the first part of $\psi \vee \psi$ has the interpretation which is different from the second part's one, the whole formula yields false.

2.2 Predicate Logic for Anaphora Resolution Language

PLA (Predicate logic for anaphora) resolution language includes the following entities:

1. Relational constants R^n
2. Individual constants $c \in C$
3. Variables $x \in V$
4. Pronoun variables p_i
5. Terms $t \in (c, x, p_i)$
6. Formulas $\phi \in (R_n t_1 ... t_n, \neg x, \exists x\phi, \phi \wedge \phi)$

Constants and *variables* are being interpreted according to the classic first-order logic. *Pronoun* p_i interpretation depends on the context.

PLA pronouns represent functions choosing the correct antecedent. Antecedent candidates are represented by the list of *terms* where i-th pronoun selects i-th existence quantifier.

2.3 Anaphora Resolution Process By Examples

Let us introduce a couple of examples to show how the anaphora is resolved. Here is the first one:

> A Kid Walks down the park ($\exists x(Kx \wedge Wx)$).
>
> There is also a Dog ($\exists y Dy$).
>
> It Frightens him and he Chases it ($Fp_1p_2 \wedge Cp_2p_1$).

Sentences are transformed into the following formula:

$$(\exists x(Kx \wedge Wx) \wedge \exists y Dy) \wedge (Fp_1p_2 \wedge Cp_2p_1) \quad (3)$$

or (in reduced form):

$$\exists y \exists x(((Kx \wedge Wx) \wedge Dy) \wedge (Fyx \wedge Cxy)) \quad (4)$$

where: Kx being the term "a Kid exists", Wx being the term "x Walks", Dy being the term "a Dog exists", Fp_1p_2 being the term "p_1 Frightens p_2", Cp_2p_1 being the term "p_2 Chases p_1".

The equivalence can be easily proved. The formula (3) requires y and x to be d ($d$og), and k ($k$id) to $Cx, Wx, Dy, Fp_1p_2, Cp_2p_1$ to be true. Thus, dk is antecedent queue.

The formula (4) produces the same result: first, the system searches the value of x such as both Kx and Wx are true. Further, y values are being searched to get Dx true. Then the final queue is dk.

The final queue fulfills the first proposition: the kid x frightens the dog y and the kid x chases the dog y. If the formula yields true, the queue dk is a correct antecedent queue.

Here is one more complex example:

> Once there was a Queen ($\exists x Qx$).
>
> Her Son Fell in Love with a frog ($\exists y(Sy \wedge \exists z(Fz \wedge Lyz))$).
>
> The prince Kissed it and she got Mad ($Kp_1p_2 \wedge Mp_3$).

The first proposition is true for every queen q. The second one is true for every son (i.e. prince) s who fell in love with a frog f. The resulting pair is sf. The queue sfq is generated after processing the first two propositions. The resulting queue is inserted into the formula so as the last proposition looks like the following: $Ksf \wedge Mq$.

Hence, the whole transformation is as follows:

$$((\exists x Qx \wedge \exists y(Sy \wedge \exists z(Fz \wedge Lyz))) \wedge (Kp_1p_2 \wedge Mp_3)) \leftrightarrow \exists y \exists z \exists x((Qx \wedge (Sy \wedge (Fz \wedge Lyz))) \wedge (Kyz \wedge Mx))$$

3. ANAPHORA RESOLVING ALGORITHM AND ITS IMPLEMENTATION

3.1 Transformation Algorithm

THEOREM 3. *If $\hat{x}$ is a sequence $x_1 ... x_n$, ϕ and ψ is closed, ϕ doesn't contain $\hat{y}$, and ψ doesn't contain $\hat{x}$:*

$$(\exists \hat{x}\phi \wedge \exists \hat{y}\psi) = \exists \hat{y} \exists \hat{x}(\phi \wedge [\hat{x}/p_i]\psi)$$

where variables x_i are free for p_i insertions in ψ

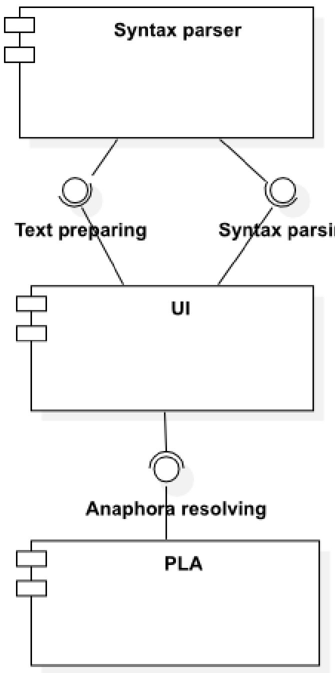

Figure 1: System structure

The formulas ϕ and ψ are closed. Also they don't contain "active" existence quantifiers and locally-resolved pronouns.

Theorem (3) shows that $\exists\hat{x}$ from the left part affects on the right part, if pronouns p_i replaced by x_i are also quantified. After $\exists\hat{x}$ coverage has enlarged, $\exists\hat{y}$ action state has to be checked, so $\exists\hat{y}$ coverage is also enlarged. The nested formulas ϕ and ψ must be closed before resolving the pronouns.

If the number of pronouns is more than of existence quantifiers, the pronoun selects an antecedent from the previous proposition.

Theorem (3) can be used for converting a PLA formula to a classic first-order logic formula. The algorithm returns a first-order logic formula $\dot{\phi}$ by using the function $[\phi]^?$ returning the effect $[\psi]^!$ on ψ by ϕ. The algorithm uses the rules as presented in definition 1:

Definition 1.

$$[(\phi \wedge \psi)]^? \rightarrow ([\phi]^? \wedge [\psi]^?);$$
$$([\exists\hat{x}\phi]^! \wedge [\exists\hat{x}\psi]^!) \rightarrow [\exists\hat{y}\exists\hat{x}(\phi \wedge [\hat{x}/p_i]\psi)]^!$$
$$[\exists x\phi]^? \rightarrow \exists x[\phi]^?; \exists x[\phi]^! \rightarrow [\exists x\phi]^!$$
$$[\neg\phi]^? \rightarrow \neg[\phi]^?; \neg[\phi]^! \rightarrow [\neg\phi]^!$$
$$[Rt_1...t_m]^? \rightarrow [Rt_1...t_m]^!$$

3.2 Implementation for Russian Texts

The algorithm described in the above section gets a syntactically processed text as input. System component structure is presented in Figure 1. After the PLA subsystem has got syntax trees for the processed sentences (see Figure 2), building a DPL becomes possible.

To proceed with the input text syntax analysis we use the component described in our earlier work [4]. For morphological analysis of the Russian texts the *Mystem*[1] analyzer

[1]https://tech.yandex.ru/mystem/

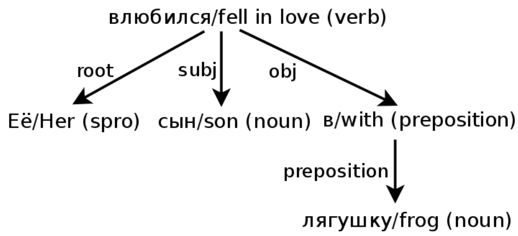

Figure 2: Syntax tree example

is used. Model required for the *MaltParser* [11] is prepared by using the National Corpus of Russian Language[2].

We implemented the user interface in two forms: CLI and a Java web application developed by using the *Play*[3].

Figure 3 represents the user web interface and illustrates how the Russian text equivalent to the above studied *queen-prince-frog* example is processed and resolved.

3.3 Anaphora Resolution Subsystem

The anaphora resolution component is implemented as an independent subsystem (which could be deployed on a separate server) communicating via *Apache Thrift*. This subsystem gets *CoNLL-X* [1] syntactically coded sentences and transforms them to the tree suitable to be represented as a DPL formula, in which the nodes are words, and the edges are their syntactic dependencies. Figure 2 shows an example of the syntax tree generated for the above mentioned sentence *"Her son fell in love with a frog"*. In the process of building a DPL formula both an object noun (e.g. *"a frog"* in our example) and a subject noun (e.g. *"son"*) are translated into the existence quantified variable with the terms $\exists ySy$ and $\exists zFz$ respectively. Object and subject pronouns are translated into the pronoun variables p_i. The predicates are represented by the terms with the above introduced quantified variables (e.g. "y fell in love with z" to Lyz).

The output data array is as follows: DPL formula, classic first-order formula, resolved sentence.

3.4 Experiments

To arrange the experiments, we used three sets of manually tagged sentences selected from the following sources:

1. 60 sentences from *Syntagrus*;
2. 60 sentences from *"Monday Begins on Saturday"*, the novel by Boris and Arcady Strugatsky;
3. 60 light-syntax structure sentences similar to the above mentioned examples.

As a result of applying the DPL based approach for anaphora resolution in Russian language texts we get the following accuracy values: *Syntagrus* – $9,6\%$; *"Monday Begins on Saturday"* – $6,4\%$; Light-syntax structures – $44,8\%$.

4. CONCLUSION

Despite the general appropriateness of the DPL to the problem of anaphora resolution, the results we achieved for the Russian texts are rather discouraging. That's why we

[2]http://ruscorpora.ru/en/

[3]https://www.playframework.com/

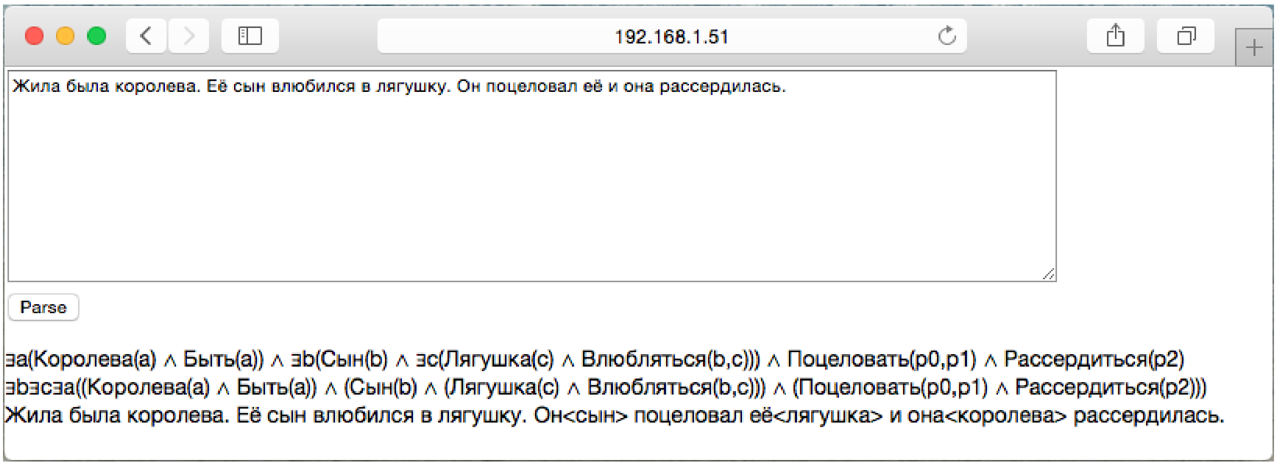

Figure 3: PLA resolution component web interface

have to conclude that both the algorithm and its implementation require further study in order to pay more attention to obvious particularities of Russian language.

Leastwise, we have to mention the following major drawbacks:

1. The model doesn't fit perfectly the task of anaphora resolution within the context of Russian text processing: many important structural linguistics properties are missing e.g. grammar case, gender, adjectives, adverbs;

2. The computational complexity of the process of building DPL formulas will be increasing significantly if we take into account more (currently missing) semantic and grammatical information which could be potentially presented in syntax trees.

To sum up, in its current state the DPL model seems to have no advantages against other known techniques in order to describe adequately the Russian language structures and to achieve satisfactory levels of anaphora resolution accuracy in Russian texts. Word position and appearence order only don't serve anaphora resolution process well. Possible efforts to take into consideration such categories as genre, number, grammatical case, object properties expressed by adjectives and adverbs, etc., might lead to higher complexity of the model without guarantee of better anaphora resolution results.

At present time machine learning techniques like [8] gain rather considerable results and look interesting for further researches and using on practice.

5. REFERENCES

[1] S. Buchholz and E. Marsi. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics, 2006.

[2] P. Dekker. *Dynamic Semantics*. Studies in Linguistics and Philosophy. Springer Netherlands, 2012.

[3] A. Ekbal, S. Saha, O. Uryupina, and M. Poesio. Multiobjective simulated annealing based approach for feature selection in anaphora resolution. In *Proceedings of the 8th International Conference on Anaphora Processing and Applications*, DAARC'11, pages 47–58, Berlin, Heidelberg, 2011. Springer-Verlag.

[4] N. Gerasimov, M. Mozgovoy, and A. Lagunov. Semantic sentence structure search engine. In *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, September 7-10, 2014.*, pages 255–259, 2014.

[5] J. Groenendijk and M. Stokhof. Dynamic predicate logic. *Linguistics and Philosophy*, 14(1):39–100, 1991.

[6] K. Karthikeyan and V. Karthikeyani. Understanding text using anaphora resolution. In *Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on*, pages 346–350, Feb 2013.

[7] V. Klyuev and Y. Haralambous. A query expansion technique using the ewc semantic relatedness measure. *Informatica: An International Journal of Computing and Informatics*, 35(4):401–406, 2011.

[8] A. Kutuzov and M. Ionov. The impact of morphology processing quality on automated anaphora resolution for Russian. In *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*, pages 232–241. Moscow, RGGU, 2014.

[9] M. Malkovskiy, A. Starostin, and I. Shilov. Method of pronoun anaphora resolution in parallel with syntactic analysis. *Perspective innovations in science, education, production and transport proceedings*, 11(4):41–49, 2013.

[10] R. Mitkov and W. W. Sb. Anaphora Resolution: The State Of The Art. Technical report, 1999.

[11] J. Nivre, J. Hall, and J. Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, volume 6, pages 2216–2219, 2006.

[12] E. Pyshkin and A. Kuznetsov. Approaches for web search user interfaces. *Journal of Convergence*, 1(1), 2010.