COREFERENCE RESOLUTION FOR UKRAINIAN TEXTS OF ECONOMIC DOMAIN

Tereshchenko V. I.
National technical university "Kharkiv polytechnic institute"
Kharkiv, Pushkinskaya str., 79/1, phone: 707-63-60
e-mail: vitalik_tereshen@mail.ru

Within the field of Natural Language Processing (NLP), the process of identifying references in the text to some entities in the same text is referred to as coreference resolution. During the recent years there has been substantial work on the important problem of coreference resolution, most of which was concentrated on the development of new models and algorithmic techniques. And despite this fact the task of coreference resolution isn`t entirely solved now.

Nowadays, the problem of coreference resolution is one of the most important problems in natural language text processing, especially when we talk about flexional languages with well-developed syntax and morphology like Ukrainian.

It is very common in Ukrainian and any other language, when people talk or write about things, such as other people, objects, or events, they use names e.g., *Барак Обама*, different kinds of descriptive expressions, e.g., *Президент*, or pronouns such as *він* or *вони*:

"Книга лежить на столі. Вона важка"

In particular, they do not have to use the same name or descriptive expression every time they mention the same thing; by choosing different ways to mention a referent, author can add more information, or stress a particular property of the person or object in question. It is unnecessary to use full names, e.g., *Олександр Терещенко*, but one can substitute these for a shorter name variant such as *Терещенко* or even *він*. And the question is how to get an idea that this or that pronoun refers to an object mentioned in the previous sentence or previous part of the sentence. This is what coreferences resolution stands for. Resolution of such things will give us more information about the text and what is it about. That is why coreference resolution is considered to be one of the most important problems in modern linguistic science.

And here is the problem. The thing is that nowadays the linguistics science is getting well-developed science and its main object is the text analysis. And it is not a secret that lately a lot of systems for textual information analysis started to appear. In order to get good results while performing natural language text analysis, one ought to be capable to resolve coreferences [5]. For this purpose a lot of different methods and approaches were developed recently and all of them works more or less good with English, but for the languages like Ukrainian there is no perfect solution of this problem. The reason for that is the fact that Ukrainian is flexional language, and it is very difficult to formalize [4].

The subject of this research is to develop an algorithms and methods applying which it is possible to resolve the problem of coreferential relations. While the

research, the following algorithms were analyzed and described: Hobbs algorithm [2], Machine Learning algorithms and also an algorithm of restrictions [1]. These algorithms were chosen to be analyzed due to the fact that the performance of their implementation usually better than the result of other algorithms.

On the basis of these algorithms analysis and taking into account the peculiarities of Ukrainian language, an algorithm for coreference resolution for Ukrainian texts of economic domain was developed. In order to define the best way for allocating coreferences in the texts, approximately100 texts of economic domain were analyzed. Within these texts coreferential relations were labeled and analyzed common factors on the basis of which an algorithm could find coreferences in the texts. [3] For example:

"…Ще в листопаді 2012 року ця американська [фірма] вільно завезла до України партію комп'ютерів. Внаслідок чого [вона] отримала прибуток що перевищив сподівання…"

In this particular case an anaphor and antecedent are in the same case, gender and number. So, we can define a rule for the algorithm: "if antecedent and anaphor have the same case, gender and number, then most likely that they refer to the same object. And this rule is true for most of the cases. But this is only one possible rule to implement into an algorithm.

As a result of the research, an algorithm for allocating coreferences in Ukrainian texts of economic domain was developed. It includes 12 rules for the purpose of precise coreferences labeling and will be implemented into the program for coreferences resolution during further research.

Taking into account the statements listed above, the step-by-step algorithm for anaphoric coreferences resolution for Russian texts of economic domain will be the following. On the preprocessing stage the pronouns and lexemes would be allocated. On the second stage the antecedents would be picked, on the basis of their morphological characteristics. The third stage is simply labeling the anaphors and their antecedents with some specific labels.

As a part of the research different types of coreferential relations, problems of natural language processing, methods of syntactic parsing and phenomena in natural language that might cause problems while text analysis were also analyzed and reviewed.

Reference list

1. *Clark J.H., Gonzalez-Brenes J.P.* "Coreference Resolution: Current Trends and Future Directions", 2008. – 11-16p.

2. *Hobbs J. R.* "Pronoun resolution" – California, 1976. – 10-18p.

3. *Ермаков А.Е.* "Референция обозначений персон и организаций в русскоязычных текстах СМИ: эмпирические закономерности для компьютерного анализа" – труды Международной конференции Диалог'2005. – Москва, Наука, 2005;

4. *Скатов Д.* "Разрешение кореференции: обзорная экскурсия" – Н. Новгород, ДИКТУМ, 2012.

5. Grosz B.J. "Readings in natural language processing" – California, Morgan Kaufmann Publishers, Inc., 1986. - p. 339-352.