УДК 681.3

O.O. Marchenko
Taras Shevchenko National University of Kyiv
Ukraine, 03680, Kyiv, Glushkova Ave., 4-d

Semantic Modification of the Mitkov Algorithm for Anaphora Resolution

О.О. Марченко
Київський національний університет імені Тараса Шевченка
Україна, 03680, м. Київ, просп. Глушкова, 4-д

Семантична модифікація алгоритму Міткова для розв'язання анафор

А.А. Марченко
Киевский национальный университет имени Тараса Шевченко
Украина, 03680, г. Киев, просп. Глушкова 4-д

Семантическая модификация алгоритма Миткова для решения анафор

The article is dedicated to modern algorithm of pronominal anaphora resolution. Anaphora resolution should be considered in a wider range of problems related with language ambiguity resolution, for instance: entity recognition, reference analysis and in general case, of course, semantic analysis of natural language text. We can render conclusion from stated above that anaphora resolution is possible only on semantic level of natural language analysis. The main purpose of this work is development of semantic heuristics for finding the most probable antecedent corresponding to anaphora with analysis of sentence context. The proposed algorithm gives about 5% improvements in comparison to the standard Mitkov algorithm.
Key Words: natural language text processing, anaphora resolution, semantic analysis.

Робота присвячена аналізу алгоритму розв'язання займенникової анафори. Розв'язання анафори має бути розглянуто в рамках широкого кола проблем лінгвістичної неоднозначності, наприклад: розпізнавання сутностей тексту, аналіз посилань та, в загальному випадку, семантичний аналіз текстів природною мовою. Із зазначеного вище можна зробити висновок, що розв'язання анафори можливе лише на семантичному рівні аналізу природної мови. Головною метою цієї роботи є розробка семантичної евристики для пошуку найбільш імовірного антецедента, що відповідає анафорі, із застосуванням аналізу контексту речень. Запропонований алгоритм дає покращення близько 5% порівняно зі стандартним алгоритмом Міткова.
Ключові слова: обробка текстів природною мовою, розв'язання анафори, семантичний аналіз.

Работа посвящена анализу алгоритма решения местоименной анафоры. Решение анафоры должно быть рассмотрено в рамках широкого круга проблем лингвистической неоднозначности, например: распознавание сущностей текста, анализ ссылок и, в общем случае, семантический анализ текстов на естественном языке. Из указанного выше можно сделать вывод, что решение анафоры возможно только на семантическом уровне анализа естественного языка. Главной целью этой работы является разработка семантической эвристики для поиска наиболее вероятного антецедента, соответствующего анафоре, с использованием анализа контекста предложений. Предложенная модификация алгоритма дает улучшение около 5% по сравнению со стандартным алгоритмом Миткова.
Ключевые слова: обработка текстов на естественном языке, решение анафоры, семантический анализ.

Introduction

Existing *rule-based* algorithms for anaphora resolution based on analysis of syntax properties have already reached their limit in quality aspects. Statistics shows that probability of connecting right antecedent with anaphora is about 85-90% [1], [2]. Further optimization is complicated by conflicts that occur due to big number of syntax rules. Attempts to optimize coefficients that determine rules priorities (so that no conflicts will occur) lead to decrease of probability for correct anaphora resolution [3]. Possible progress is not significant (within 0.1-0.001%).

Anaphora reference resolution is impossible without semantics of candidates to antecedent and analysis of how its semantic meaning consistent with semantic of words that close neighbor to anaphora. As it has been stated before, the main purpose of this article is creation of semantic heuristics for correct antecedent determination with usage of semantic meaning of sentences. We decided to do it with modification of existing methods through adding semantic rules into the Mitkov algorithm [2].

There are wide arrays of approaches for solving this problem. The modern approaches are: Lapin and Leass algorithm, centering algorithm, Hobb's algorithm, Mitkov algorithm [1], [2]. Mitkov algorithm is known to be quite flexible and adjustable. Ordinary realization of Mitkov algorithm doesn't solve dubious cases and can cause wrong answer. Semantic rules like "semantic triplet match" together with syntax restrictions can give us improvement in ambiguous cases. This helps to extend a "bottleneck" of Mitkov algorithm.

In the next sections, we discuss our new resolution algorithm and statistics about some resolution.

Mitkov algorithm

Mitkov algorithm to pronominal anaphora can be described as a set of rules that weight candidates to antecedents and after that the best candidate is antecedent with greatest salience. The set of rules is:

1. **Definiteness**: All defined nouns have weights +1, candidates that don't have any defined nouns : -1;
2. **Giveness**: candidates that represent the following topic: +1;
3. **Indicator words**: candidates after verbs: {discuss, present, illustrate, identify, summarize, examine, describe, define} have +1;
4. **Lexical reiteration**: repeated candidates have salience +1 if they are repeated once, +2 if they are repeated twice, and so on;
5. **Non-pronominal phrases**: candidates that enter NP have salience +1;
6. **Collocation pattern preference**: +2 to salience of candidates that have syntactic position the same that a pronoun;
7. **Connective pattern**: in case like: "you V1 NP or con((you) V2 it) con ((you) V3 it)" NP candidates have +2 to it's salience;
8. **Reminder indicator**: candidates that are reminded in previous sentences have +1, in the same sentence : +2;
9. **Field indicator**: candidates that concern the same field that antecedent have +1;
10. **Boost pronoun**: candidates that have more references to pronouns have more salience;
11. **Syntax parallelism**: in case of same syntactic position, candidates have +1;
12. **Reference indicator**: in case of the most referred antecedents, they have +1.

Analysis and improvements of the Mitkov algorithm's

The Mitkov algorithm is a rule-based approach. It is based on syntax rules that can conflict with each other. The main "bottleneck" in this algorithm in some relations must be solved on semantic level but this can not be done, because we have only syntax-based rules. In this case probability of finding right antecedent is very low. This problem can be solved only with semantic rules implementation. If we try to extend the set of syntax rules this could only increase conflicts between the rules and will lead to wrong antecedent as a result. We can also strengthen pronominal anaphora algorithm with semantic measurement implementation. The main advantage of the Mitkov algorithm is that it can be easily adjusted so that we can process a set of sentences. In our realization we look backwards for four sentences.

Let us consider rule #6: Collocation match pattern. We can extend this rule adding some semantic sense to it. We can increase probability of choosing right antecedent modifying this indicator. For instance one of possible modifications can be done: we can see that semantic position of candidate is not strictly the same as semantic position of pronoun but close enough. Close enough means that semantic distance less or equal than value that was specified before. In our case we used semantic distance by Leacock and Chodorow: $sim_{lch}(c_1, c_2) = -\log\left[\frac{len(c1, c2)}{2 * MAX}\right]$, where *len* is the number of edges on the shortest path in the taxonomy between the two concepts(words) and MAX is the depth of the taxonomy [4], [5].

Another possible modification is to create triplets in a form like: VERB *verb* NOUN *noun1* NOUN *noun2*. With this triplet we can use semantic distance for *noun1* and *noun2*. There can be more modifications done. They are building layer by layer so that chances of finding right antecedent will raise a lot.

We can also use syntactic restrictions when composing weights for antecedents - NPs.

Let us consider the following example:

"To avoid data loss on devices, we should avoid storage of critical information on them".

Let's take a look for possible outcome of our algorithm using the Mitkov approach without any syntactic restrictions: $storage\begin{pmatrix} data, 1 \\ devices, 2 \\ losses, 3 \end{pmatrix}$

as we can see that our most probable antecedent is "data". But if we use restriction like "Anaphora can not refer on co-argument" [6], [7] then algorithm cut off "data" case, and we can have "devices" as the most probable antecedent.

We can use syntactic restrictions like:

Pronoun P and noun phrase N are non-coreferential if any of the following conditions are hold:

1. P and N have incompatible agreement features.
2. P is in the arguments domain of N.
3. P is the adjunct domain of N.
4. P is an argument of head noun, N is not pronoun, and N is contained in head noun.
5. P is in the NP domain of N.
6. P is determiner of a noun Q, and N is contained in Q.

Main idea

The main concept is eliciting semantic information that concern anaphora and try to find noun with a similar semantic context. In our algorithm we look backwards, up to five-

six sentences, and try to find closest semantic triplet. In cases when antecedent is a pronoun we can substitute anaphora that we are resolving with that pronoun and find antecedent for new pronoun. When we will have a situation that antecedent is noun, we can trace back to our original pronoun.

For instance let us consider example:

"John likes to solve extraordinary problems. Last set seams to be quite difficult, it took almost all day him to solve them. His extraordinary talent gives him advantage, that's why he is obsessed in solving them".

Let us try to solve anaphora for "them":

First triplet: VERB(solving) NOUN1(them) NOUN2(he)
Seconds triplet: VERB(took) NOUN1(his) NOUN2(them)
Third triplet: VERB(solve)NOUN1(John)NOUN2(problems)

Together with semantic approach, pronoun substitution we can come that "he" refers to "John" and "them" – to "problems".

Let us consider another example:

"There was seen a tail of a fox. It has stolen a chicken. It was red, furry and with a white tip".

In this case syntax structure is identical, and it's impossible to determine with only syntax rules antecedent for last "it". With usage of semantic rules we can determine that "tail" can not steal something, so first "it" is reference to "fox". Second "it" cannot be reference to a fox, because "tip" is not a property of a fox but of a tail. So second "it" will correspond to "tail".

Experiments

Here are some statistics demonstrated improvements of a new algorithm over the Mitkov standard.

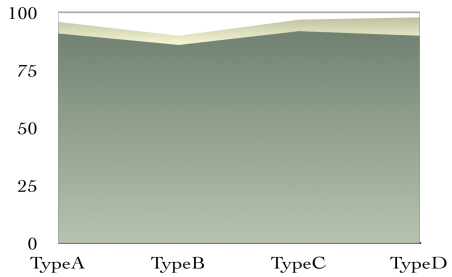

Fig. 1 – Improvements of the New Algorithm

On the graphic above the statistical information is presented: dark area is results of the Mitkov algorithm without any modifications, light area is the results of the Mitkov algorithm with modifications (semantic measures, semantic approach) on different types of sentences. On Y axis, there is probability that antecedent correctly matched with pronoun. On X axis, we can see the sets of sentences that were used to test and compare the improved Mitkov algorithm and the standard Mitkov algorithm.

Type A sentences is complex sentences that can have more that one NP in them. Type B is sequence of sentences, maximum length is two sentences. Type C is the sequence where maximum number of sentences is three, Type D is the sequence with the length of four sentences. On each type, there were about ten sentences.

As we can see, modification gives us improvement about 4-5%, on every type of a sentence.

Conclusions

Analysis of existing algorithms has convinced us that it's impossible to solve anaphora resolution problem just by syntax means. Semantic rules are needed so that it's possible to determine which of candidates to antecedents is closer to words – context neighbours of anaphora. Semantic rules were created so and added to the modern Mitkov algorithm.

In current realization we were able to improve rule-based algorithm and we have provided evidence with statistical data. Using of semantic approach allows us to solve some cases when in syntax approach we have to use priority of rules that was determined in empirical way. With usage of semantic rules we can determine antecedent more precisely.

Usage of semantic distance metrics that have been constructed on the base of global ontology networks can give some improvements to procedure of context linkage of candidate to antecedent to anaphora place.

References

1. Carbonell J.G. Anaphora: analysis, algorithms and applications / J.G. Carbonell, J. Siekmann. – Springer , 2007. – P. 125-150.
2. Antonio Branco. Anaphora processing. Linguistic, cognitive and computational modelling (Book style) / Antonio Branco Tony McEnergy, Ruslan Mitkov. – 2005. – ch 2, 3.
3. Chierchia G. Dynamic of Meaning: Anaphora, Presupposition and the Theory of Grammar / G. Chierchia. – University of Chicago Press, 1995. – ch. 4.
4. Leacock C. Using Corpus Statistics and WordNet Relations for Sense Identification, Computational Linguistics - Special issue on word sense disambiguation Claudia Leacock, George A. Miller , Martin Chodorow – March 1998. – Vol. 24, Is.1. – P. 147-165.
5. Budanitsky Al. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures, Workshop on wordnet and other lexical resources / Alexander Budanitsky, Graeme Hirst. – 2001.
6. Mitkov R. A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method. CICLing '02 / Ruslan Mitkov, Richard Evans, and Constantin Orasan // Proc. of the Third International Conference on Computational Linguistics and Intelligent Text Processing. – 2002. P. 168-186.
7. Patrick Sturt A New Look at the Syntax-Discourse Interface: The Use of Binding Principles in Sentence Processing, Journal of psycholinguistic research,Volume 32, Number 2 (2003), pp.125-139.

RESUME

O.O. Marchenko

Semantic Modification of the Mitkov Algorithm for Anaphora Resolution

The article is dedicated to modern algorithm of pronominal anaphora resolution. Anaphora resolution should be considered in a wider range of problems related with language ambiguity resolution, for instance: entity resolution, reference analysis and in general case, of course, semantic analysis of natural language text. Analysis of existing algorithms has convinced us that it's impossible to solve anaphora resolution problem just by syntax means. We can render conclusion from stated above that anaphora resolution is possible only on semantic level of natural language analysis. The main purpose of this work is development of semantic heuristics for finding the most probable antecedent corresponding to anaphora with analysis of sentence context.

Semantic rules are needed so that it's possible to determine which of candidates to antecedents is closer to words - context neighbours of anaphora. Semantic rules were created so and added to modern Mitkov algorithm.

In current realization we were able to improve rule-based algorithm and we have provided evidence with statistical data. Using semantic approach allows us to solve some cases when in syntax approach we have to use priority of rules that was determined in empirical way. With usage of semantic rules we can determine antecedent more precisely.

Usage of semantic distance metrics that have been constructed on the base of global ontology networks can give some improvements to procedure of context linkage of candidate to antecedent to anaphora place. The proposed algorithm gives about 5% improvements in comparison to the standard Mitkov algorithm.

Статья поступила в редакцию 30.05.2012.