ОПЫТ АНАФОРИЧЕСКОЙ РАЗМЕТКИ КОРПУСА И РАЗРЕШЕНИЯ АНАФОРЫ С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ

Протопопова Е. В. (protoev@gmail.com),
Бодрова А. А. (anastasie.bodrova@gmail.com),
Вольская С. А. (svetlana.volskaya@gmail.com),
Крылова И. В. (krylova93@gmail.com),
Чучунков А. С. (scarywound@gmail.com)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Алексеева С. В. (alexeeva@opencorpora.org),
Бочаров В. В. (bocharov@opencorpora.org),
Грановский Д. В. (granovsky@opencorpora.org)

Проект «Открытый Корпус», OpenCorpora.org

Ключевые слова: корпуса текстов, краудсорсинг, разрешение анафоры, синтаксическая разметка

ANAPHORIC ANNOTATION AND CORPUS-BASED ANAPHORA RESOLUTION: AN EXPERIMENT

Protopopova E. V. (protoev@gmail.com),
Bodrova A. A. (anastasie.bodrova@gmail.com),
Volskaya S. A. (svetlana.volskaya@gmail.com),
Krylova I. V. (krylova93@gmail.com),
Chuchunkov A. S. (scarywound@gmail.com)

Saint Petersburg State University, St. Petersburg,Russia

Alexeeva S. V. (alexeeva@opencorpora.org),
Bocharov V. V. (bocharov@opencorpora.org),
Granovsky D. V. (granovsky@opencorpora.org)

OpenCorpora.org

The paper describes the noun phase and anaphora annotation in OpenCorpora and compares it to that in other corpora. We discuss the choice of representative texts for anaphoric annotation and the basic principles of syntactic annotation. In case of noun phrase annotation we followed the scheme introduced earlier for morphological annotation: it was carried out in two stages: firstly, all noun phrases and some other syntactic units were annotated by a heterogenous group of people, then a linguist compared all markup results and found the best one, or corrected mistakes. We present some annotation results and cases of annotator's disagreement and proceed to introduce our data-driven anaphora resolution system based on decision trees. We then list the features used to fit the classificator and discuss their relevance and some changes which improved the classificator performance. We also present out rule-based approach to automated noun phrase extraction using Tomita parser. A baseline for anaphora resolution is introduced and we compare it with our results.

Keywords: anaphora resolution, corpora, crowdsourcing, syntactic annotation

1. Introduction

The task of anaphora and coreference resolution is quite important in automated text processing and understanding, since these are often used to maintain text coherence. Facing this task state-of-the-art coreference resolution systems exploit various supervised machine learning techniques [13] and thus require manually annotated data. Our goal is therefore to build such a corpus for Russian as a part of OpenCorpora project[1]. Moreover, no such corpus for Russian is freely available now and that is perhaps why the number of papers on using machine learning techniques for Russian coreference resolution is rather low.

We start with annotating noun phrases in a part of our corpus including some news articles and short stories. Groups of words acting as one syntactic unit, which contain nouns, (such as complex preposition *в связи с 'in view of'*) were also annotated. We tried to make our annotation scheme as simple as possible because some annotators are not linguists. A small portion of texts was annotated by several people, the others being given only to one person. A linguist then compared (in case of several annotations) or corrected the resulting annotation. Then a text was annotated with anaphoric relations for 3rd person pronouns, some demonstrative pronouns in anaphoric use, possessive pronouns and relative pronouns (*который 'which / that'*, *кто 'who'* etc.). An annotator can only mark the relation between a selected pronoun and previously annotated NP, this is why the previous stage is important. We later discuss the problem concerned with choosing the right phrase in case there are several referring to one entity.

One of the ways to evaluate such a corpus is to use it for the task of anaphora resolution. We build a classifier which uses some morphological and textual features. The corpus annotated for anaphora resolution track was used as training data. A tool for automated noun phrase extraction was implemented using Tomita-parser[2]. With its help positive and negative training examples were then generated.

[1] OpenCorpora, available at: http://opencorpora.org/

[2] Tomita-parser, available at: http://api.yandex.ru/tomita/

2. Related Work

Two tasks may be distinguished in our work: first of all, we describe our experience in anaphora annotation, then, we present the anaphora resolution system. It seems natural to mentions previous work made in these two fields in two separate subsections.

2.1. Anaphoric Annotation in Corpus

A great number of annotation schemes were proposed for anaphoric and coreferential annotation. The most famous is MUC scheme[3] which aims at high inter-annotator agreement and was used in MUC competitions. The annotation rules are quite simple; personal and possessive pronouns, names and other named entities, bare nouns as a part of coreference chain are markables (should be annotated). As for relations, they annotate basic coreference (two NP refer to the same object), bound anaphors, apposition and some other more specific cases. Another significant resource for English was developed as a part of Ontonotes corpus [2]. An important difference from MUC (ACE) scheme concerns the annotation of verbal phrases: in Ontonotes such phrases may be marked as antecedents or coreferring phrases (There is an example of predicative anaphora in section 3.3). Anaphoric and bridging (or associative) relations were also annotated in the ARRAU corpus [9], [10]. In general, they followed MUC scheme, but propose another markup format based on TEI instructions[4].

As for Slavonic languages, an annotation scheme different from previous ones was proposed for Prague Dependency Treebank (PDT) [7]. All referential entities including generic and abstract ones were subject to annotation. Predicate nominals and appositions were not considered as corefent. Textual coreference including pronominal one is annotated along with bridging relations.

2.2. Anaphora Resolution Using Machine Learning Techniques

Many anaphora resolution systems proposed in last 50 years are rule-based. They are, however, worth concerning because their rules use so called anaphora resolution factors. Perhaps an exhaustive survey of such systems is presented in [6]. A recent highly appreciated system is Stanford Coreference Resolution system [12]. We will now focus on some important works regarding these factors as well as machine learning approach to anaphora resolution.

[13] propose a corpus-based system, which learns from small amount of data using quite a restricted number of features. The vector for each pair of markables consists of the following 12 factors: distance in sentences, whether each markable is a pronoun, whether markables are equal, whether NP is definite or demonstrative,

[3] MUC-7 Coreference Task Definition, available at: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html

[4] TEI:Text Encoding Initiative, available at: http://www.tei-c.org/index.xml

agreement in number, gender and semantic class (each counted separately), alias (i.e. if NP is another name for the same entity) and appositive features. C5 learning algorithm (based on decision tree) was used to learn from this data. The evaluation was conducted on MUC sets, the best result reported achieved F-measure of 62.6% and precision 58.6%. They also analyzed classification errors. The improvements to this system were proposed in [8], which lead to F-measure of 70.4%. They used 53 features but then reduced this number to 41, including various syntactic and semantic features.

Different machine learning techniques were tested in [5].Using previously examined factors of referential choice [3], [4], they achieved quite a high accuracy of anaphora resolution up to 88.7%.

3. Corpus Annotation

3.1. Texts for Annotation

Our chief aim was to create a representative corpus, which may be then used as a training data for anaphora resolution systems, so texts for annotation were chosen with special attention. First of all, we considered genre structure of this subcorpus. We consider press materials to be the most characteristic of modern language and in particular of such phenomena as anaphora and coreference, that is why a half of our subcorpus is composed of news articles. Another half is made up of fiction texts, blog posts and encyclopedic texts. Then the texts were filtered automatically: we examine news size and choose texts of its average size for all listed genres. Then they were reviewed manually and were filtered again. We exclude texts which do not include many examples of anaphoric relations. About a third of texts were filtered out and were substituted by other texts with the highest number of pronouns.

The total size of corpus planned was about 100,000 words. By the time, however, only 18,000 are annotated.

3.2. Noun Phrase Annotation

As mentioned above, we started with annotating noun phrases and some more specific units. The following kinds of phrases are subject to annotation: basic noun phrases, names and named entities, pronouns, complex conjunctions and prepositions, parenthetical expressions and prepositional phrases. The exhaustive list of groups is presented in table 1.

Each annotator chooses a text and annotate it sentence by sentence. The annotation process may be divided into the following steps: the annotator first finds all nouns in the phrase and then marks all simple groups (1–7). Basic noun phrases may include adjectives, ordinal numbers, adverbs (очень 'very') and particles (не 'not'). An annotator should mark group as proper name if it contains a proper noun. Thus, in the following expression 'Марина Павловна Трубецкая' 'Marina Pavlovna Trubetskaja' three groups of the second kind should be annotated. Groups 4–7 are specified by the lists from Russian National Corpus (RNC)[5]. Complex groups (8–15) should include at least two simple groups.

Table 1

	Group	Example
1	basic noun phrase	*не очень интересный журнал*
2	proper name	*прекрасную Францию*
3	numeral	*сто двадцать пять*
4	complex preposition	*в течение*
5	adverbial expression	*без оглядки*
6	parenthetical expression	*к слову сказать*
7	complex conjunction	*до тех пор пока*
8	complex proper names	*Марина Павловна Трубецкая*
9	proper name with generic term	*княжна Трубецкая*
10	apposition	*статья 112*
11	prepositional phrase	*от меня*
12	coordinated NPs	*Маша и Петя*
13	complex noun phrase (NP containing two or more NPs)	*куртка Маши*
14	numeral phrase	*три яблока*
15	complex pronoun	*друг друга*

An annotator should also mark heads for those phrases where it is not obvious. We consider it to be obvious in cases where head can be easily found automatically: basic NPs, prepositional phrases, enumeration. We also introduce special tags ALL and NONE for enumerations and groups 4–7 respectively.

When all sentences of the text are annotated, the mark-up should be revised. A moderator reviews the annotation sentence by sentence and can accept annotated groups or mark their own.

3.3. Anaphoric Annotation

For the anaphoric annotation the pronouns from the list were highlighted in text and all annotators can mark relations between these pronouns and preceding NPs. The annotation follows several rules: first of all, we agreed to mark the relation between a pronoun and its nearest member of coreferential chain. Thus, in the following sentence, the relation between 'Фернандо Алонсо' 'Fernando Alonso' and 'свой' 'his' should be annotated (1):

(1) ***Фернандо Алонсо*** *в первый раз в* ***своей*** *карьере пилота Формулы-1 выиграл Гран-при Монако.* *'****Fernando Alonso*** *won Grand-Prix Monaco for the first time in* ***his*** *Formula 1 driver career'*

[5] http://ruscorpora.ru

Moreover, the antecedent should be the maximal possible group. We do not annotate predicative anaphora such as (2):

(2) ***Шёл дождь***. ***Это*** *нас остановило.* '***It was raining***. ***This*** *stopped us.*'

We do not annotate cataphora though we have seen several examples of it in texts such as:

> *Хотя* ***он*** *казался спящим,* ***Иван*** *думал.*
> *'Although* ***he*** *seemed to be sleeping,* ***John*** *was thinking.'*

One reason for this is that we would like to limit classifier's search space and the number of possible antecedents in text.

4. Anaphora Resolution System

We implemented a data-driven anaphora resolution system, which relies on previously annotated corpus. The pairs of markables in corpus are deduced automatically by a special tool for noun phrase extraction and the training vectors are computed for all possible pairs 'antecedent—anaphora'. Pairs are marked as positive/negative examples and then are used to fit the classifier. These stages are described in corresponding subsections.

4.1. Noun Phrase Extraction

To extract all possible markables equivalent to those used in manual annotation we developed a NP extraction tool using Tomita-parser[6]. Originally a tool for fact extraction, Tomita deals with context-free grammars and key-word dictionaries. For the current purpose, a grammar for NP extraction was used to process sentences. For each rule, the parser finds the longest substring meeting the requirements. Thus, our groups were defined in terms of sequences of tags. Sometimes, our restrictions were insufficient and the rules were corrected many times. An XML output of parser was then combined with the information from our tokenizer and a markable was represented as a pair of identifiers—text id and token id. Precision and recall are 0.81 and 0.82 respectively.

4.2. Feature Vectors

A set of features is necessary for a classifier to define whether a pair is bound with anaphoric relation or not. Our features are based on practical as well as theoretical conclusions and are meanwhile easy to compute. All extra information was obtained through open-source tools and resources.

[6] Tomita-parser, available at: http://api.yandex.ru/tomita/

We divide our features into three groups: lineal, morphological and syntactic features. These classes are described below. Each feature was computed for anaphor and its possible antecedent head.

Lineal Features

1. The number of proper nouns between anaphor and antecedent
2. The number of sentences between anaphor and antecedent
3. The number of potential anaphors for the given antecedent between given anaphor and given antecedent
4. The number of nouns between anaphor and antecedent
5. The number of anaphoric pronouns between anaphor and antecedent
6. The number of possible antecedents for the given anaphora between given anaphor and given antecedent

Morphological features

These features were computed using our morphological mark-up (OpenCorpora morphological dictionary) and no disambiguation was carried out.

1. Part-of-speech of the antecedent
2. Whether antecedent is in nominative
3. The number of verbs in the sentence containing antecedent
4. The number of nouns in the sentence containing antecedent
5. The number of conjunctions and pronominal adjectives in the sentence containing antecedent
6. The number of nonfinite verb forms in the sentence containing antecedent

Syntactic features

The syntactic information for these features was obtained with the help of MaltParser[7].

1. Whether antecedent is subject
2. Whether anaphor is subject

4.3. Classifier

Our learning method is based on decision trees and follows the ID3 algorithm [11]. Test pairs were extracted from documents as it was preciously done for training pairs. The vectors were post-processed, because the result on the data as is was very low (18% accuracy). The following steps were undertaken:

1. Binarize all lineal features to features 'is more than', 'is between X and Y' etc.
2. Remove all pairs where no positive pair is found for a pronoun.
3. Treat all examples where antecedent is too far from anaphor as negative.
4. Add feature counts from the nearest possible antecedent to all possible antecedents for given pronoun.

[7] MaltParser, available at: http://www.maltparser.org/

The classifier starts from the anaphor and proceeds till a positive pair is found. Then it works till the first negative example and marks all further pairs as non-anaphoric.

5. Evaluation

In this section we would like to present the results of manual annotation as well as the results of automated anaphora resolution.

5.1. Manual Annotation

Although at first there were controversial opinions on the annotation principles, the annotation itself seems to be quite simple for annotators. Seven annotators participated in this task and 9,100 groups (5,788 simple and 3,312 complex) were annotated. First of all, we can observe annotator-moderator agreement. The annotators make mistakes only in difficult cases (such as the order of combining units into complex groups) though they are no professional linguists.

We use two metrics to estimate inter-annotators' agreement: Cohen's kappa [15] and F-mean. They show quite good results for pairs of annotators: for simple groups kappa varies from 0.61 to 0.97 and F-mean is more than 0.9. The results concerning complex groups are somewhat lower: best kappa scores vary from 0.67 to 0.75. These figures suggests that the annotation manual was clear for all annotators and that the task itself is not very difficult.

5.2. Anaphora Resolution

Our baseline system marks as anaphoric pair pronoun and the nearest possible antecedent. The accuracy is computed in the following way: a pair is marked correctly as anaphoric if its antecedent's head equals to that of reference antecedent. Baseline accuracy is therefore about 50.4% on the corpus of 94 documents and somewhat more than 2,000 pairs.

The current system was built using the corrections mentioned above and achieved the accuracy of 52.04%. Here we present a part of the tree (1 is for antecedent and 2 is for anaphor) (3):

```
(3)  'number of NPRO',
     {
       '>4' => [
          'POS 2',
          {
            '2_3' => 'no',
            '3' => [
                      'number of nonfinites for 2',
{
 '>10' => 'no',
 '<undef>' => 'no',
 '1' => 'no',
 '2' => [
         '1 is nominative',
         {
          '1' => [
          'number of nouns <= 3 for 1',
          {
           '1' => 'no',
           '0' => 'yes'
          }
```

6. Conclusion

In this paper we have described our attempt to create a corpus with anaphoric annotation and an anaphora resolution system. Here we would like to outline some of the future directions. First of all, we have seen that the part of NP annotation is time-consuming so it may be conducted in semi-automated way as we have already implemented a tool for fully automated NP extraction. The anaphora resolution system may be improved with many additional features and, furthermore, be transformed into coreference resolution system. On the other hand, we can pay more attention to the training data with respect to proportion of positive and negative examples and more complicated learning algorithms.

References

1. *Bocharov V. V., Alexeeva S. V., Granovsky D. V., Protopopova E. V., Stepanova M. E., Surikov A. V.* (2013) Crowdsourcing morphological annotation [Morfologicheskaja razmetka korpusa silami volontërov], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2013"], Bekasovo, pp. 109–115
2. *Hovy E., Marcus M., Palmer M., Ramshaw L., Weischedel R.* (2006), OntoNotes: The 90% Solution, available at: http://bbn.com/resources/pdf/HLT-NAACL-2006-OntoNotes.pdf
3. *Kibrik A. A., Dobrov G. B., Zalmanov D. A., Linnik A. S.* (2010), Referential choice as a multi-factor probabilistic process [Referentsial'nyj vybor kak mnogofaktornyj verojatnostnyj protsess], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2010"], Bekasovo, pp. 173–181

4. *Kibrik A. A.* (2011), Reference in Discourse, Oxford Studies in Typology and Linguistic Theory
5. *Kibrik A. A., Linnik A. S., Dobrov G. B., Khudyakova M. V.* (2012), Optimizing a machine learning base model of referential choice [Optimizatsija modeli referentsial'nogo vybora, osnovannoj na mashinnom obuchenii], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2012" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2012"], Bekasovo, pp. 237–247
6. *Mitkov R.* (1999), Anaphora resolution: the state of the art, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.6235&rep=rep1&type=pdf
7. *Nedoluzhko A., Mírovský J., Ocelák R., Pergler J.* (2009), Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank, available at: http://ufal.mff.cuni.cz/~nedoluzko/koref_anot/DAARC_Nedoluzhko.pdf
8. *Ng V., Cardie C.* (2002), Improving Machine Learning Approaches to Coreference Resolution, available at: http://acl.ldc.upenn.edu/acl2002/MAIN/pdfs/Main347.pdf
9. *Poesio M.* (2004), The MATE/GNOME Proposals for Anaphoric Annotation, Revisited, available at: http://acl.ldc.upenn.edu/W/W04/W04-2327.pdf
10. *Poesio M., Artstein R.* (2008), Anaphoric Annotation in the ARRAU Corpus, available at: http://catalog.ldc.upenn.edu/docs/LDC2013T22/lrec08_297.pdf
11. *Quinlan J. R.* (1986), Induction of Decision Trees. Machine Learning, Vol. 1, I. 1. available at: http://www.dmi.unict.it/~apulvirenti/agd/Qui86.pdf
12. *Recasens M., De Marneffe M., Potts C.* (2013), The Life and Death of Discourse Entities: Identifying Singleton Mentions. In Proceedings of NAACL-HLT 2013.
13. *Soon W. M., Ng H. T., Lim D. C. Y.* (2001), A Machine Learning Approach to Coreference Resolution of Noun Phrases, available at http://acl.ldc.upenn.edu/J/J01/J01-4004.pdf
14. *Lee H., Chang A., Peirsman Y., Chambers N., Surdeanu M., Jurafsky D.* (2013), Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. Computational Linguistics, Vol. 39, N. 4. MIT Press, Cambridge, MA.
15. *Cohen, Jacob* (1960), A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1).