ФАКТОРЫ РЕФЕРЕНЦИАЛЬНОГО ВЫБОРА: КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

Н. В. Лукашевич (louk@mail.cir.ru)

Г. Б. Добров (wslc@rambler.ru)

МГУ, Москва, Россия

А. А. Кибрик (aakibrik@gmail.com)

Институт Лингвистики, Санкт-Петербург, Россия

М. В. Худякова (mariya.kh@gmail.com)

А. С. Линник (skylinnik@gmail.com)

МГУ, Москва, Россия

Выбор между различными типами референциальных выражений, таких как дескрипции, имена собственные и местоимения, зависит от большого числа одновременно действующих факторов. В данном исследовании роль и значимость этих факторов моделируется при помощи различных алгоритмов машинного обучения. Работа основана на специальном англоязычном корпусе RefRhet, размеченном по референции.

Ключевые слова: компьютерное моделирование, референциальный выбор, референциальное выражение, факторы, RefRhet.

FACTORS OF REFERENTIAL CHOICE: COMPUTATIONAL MODELING[1]

N. V. Loukachevitch (louk@mail.cir.ru)

G. B. Dobrov (wslc@rambler.ru)

Lomonosov Moscow State University, Moscow, Russian Federation

A. A. Kibrik (aakibrik@gmail.com)

Institute of Linguistics, Russian Academy of Sciences, Moscow, Russian Federation

M. V. Khudiakova (mariya.kh@gmail.com)

Lomonosov Moscow State University, Moscow, Russian Federation

A. S. Linnik (skylinnik@gmail.com)

Lomonosov Moscow State University, Moscow, Russian Federation

Referential choice between various referential expressions, such as descriptions, proper names, and pronouns, depends on a variety of factors. We present recent results of our modeling study into referential choice, based on the RefRhet corpus. The account of additional factors and the employment of mixed machine learning techniques enabled an improvement of referential choice prediction. This applies both to the two-way choice between full NP and pronoun and to the threeway choice "descriptive full NP vs. proper name vs. pronoun". We have demonstrated that the great majority of the factors taken into account are signifi cant for modeling the referential choice.

Key words: computational modeling, referential choice, referential expression, factors, RefRhet.

[1] This study was supported by grant #09-06-00390 from the Russian Foundation for Basic Research.

Introduction

When producing discourse, speakers or writers constantly face the necessity to mention persons or objects, that is, perform reference. When performing reference, a speaker or writer chooses between several major forms of reference, including pronouns, descriptive noun phrases, or proper names. We call this procedure referential choice.

Choosing an appropriate form of reference is apparently crucial for the overall felicity of the created discourse. Modeling referential choice is, therefore, an important part of the technologies of language generation. Referential choice is also related to the technologies of automatic text summarization. For example, Nenkova (Nenkova 2008) notes that among the major difficulties of the modern technologies of summarization, based on the identification of key sentences, is the referential organization of discourse. In particular, according to the results of DUC conference (http://duc.nist.gov/) it was found that over one half of automatically generated summaries mention entities, whose relation to the reported events remains unknown.

There is extensive linguistic literature on referential choice, see e.g. (Givón 1983), (Fox 1987). During the last decades computational models of referential choice have appeared, too, see e.g. (Strube, Wolters 2000). Kibrik (Kibrik 1996, 1999) proposed a calculative model of referential choice that pinpointed a number of factors with certain numerical weights. The sum of the weights was supposed to predict referential choice between a full NP and a pronoun. Grüning and Kibrik (Grüning, Kibrik 2005) attempted a model of referential choice in which the contribution of individual factors was defined automatically, and the interaction between factors was allowed to be non-linear. This model was based on neural networks, a well-known algorithm of machine learning. All of the mentioned Kibrik's studies explored small data sets counting one or two hundred referential expressions. In fact, the use of machine learning requires much greater data sets, which presupposes the creation of large corpora annotated for reference.

One corpus of this kind was formed for the GREC conference, see e.g. http://www.nltg.brighton.ac.uk/research/genchal10/grec/. GREC participants were supposed to demonstrate automatic systems generating appropriate referential expressions for the central entity spoken of in a text. A referential corpus of 2000 Wikipedia articles, describing people, countries, landscapes, etc., was collected and annotated for reference. 90% of the corpus was provided to conference participants for training their systems, and the results were demonstrated on the basis of a test subcorpus (10% of the corpus). Participants were allowed 48 hours for providing their results on the test subcorpus.

In (Kibrik et al. 2010) we described the computational modeling of referential choice, based on the specially designed RefRhet corpus. In this paper we discuss the specifics of referential factors, only briefly mentioned in (Kibrik et al. 2010). We also present the new results of our project.

1. RefRhet corpus: The present stage

The RefRhet corpus is based on the English-language corpus RST Discourse Treebank, created under the direction of Daniel Marcu

(http://www.isi.edu/~marcu/discourse/Corpora.html), see (Carlson et al. 2003). The corpus contains 385 Wall Street Journal articles on economics and politics. These articles contain 176 383 words and 21 789 elementary discourse units. This corpus was chosen as the basis for RefRhet because it contained annotation for rhetorical structure. Rhetorical structure has been shown to be important for reference in discourse (Fox 1987); on its basis Kibrik (Kibrik 1996) proposed the measurement of rhetorical distance that proved significant in the studies of referential choice.

The notion of rhetorical distance (RhD) is based on Rhetorical Structure Theory (Mann and Thompson 1988). This theory describes the hierarchical semantic organization of discourse. Each elementary discourse unit (most typically coinciding with a clause) is a minimal node in a rhetorical net. Terminal nodes are combined into groups in accordance with hierarchical closeness. Nodes, both terminal and complex, are connected by rhetorical relations, either symmetric (sequence, conjunction, contrast) or asymmetric (cause, condition, concession, etc.). Rhetorical distance is the measurement of the path along the rhetorical net from one node to another.

Rhetorical distance between clauses helps to take into account those instances in which the anaphor unit and the antecedent unit are hierarchically close but linearly far apart, and vice versa.

Referential annotation was added to RST Discourse Treebank, and as a result the RefRhet corpus emerged. Referential annotation was performed with the help of the MMAX-2 program, created by a group of German computational linguists specifically for modeling reference (see http://mmax2.sourceforge.net/). MMAX-2 annotation is done with the help of a so-called annotation scheme (Krasavina, Chiarcos 2007). The annotation scheme employed contains a set of annotated parameters, or factors.

An element that undergoes annotation, called markable, is a text constituent that can serve as a referential expression. Coreference relations are posited between markables. A coreference relation connects any non-first mention *n* of a referent (that is, anaphor), with the previous mention *n–1* (that is, antecedent). In addition, each markable contains a number of annotated features (grammatical role, animacy, etc.) that can affect referential choice.

Since all of the annotations are performed manually, a certain number of mistakes is inevitable. In order to exclude such mistakes the decision has been made to annotate each text twice and then compare these annotations automatically. Such comparison results in a list of markables that either appear only in one of the annotations, or have different feature values in the two annotations. Subsequently, annotators from a different group choose the correct analysis out of the two available.

The present-day stage of the RefRhet corpus is as follows: 157 texts are annotated twice, 193 texts are annotated once, and 35 texts are not yet annotated. The RefRhet corpus is among the largest of its kind that exist to date; cf. (Byron and Gegg-Harrison 2004; Ge et al.) 1998; Tetrault 2001; Orasan 2004; GREC corpora. Given that the annotation of a referential corpus is an extremely laborious task, creating a larger corpus would simpler be unpractical. From the statistical point of view, the corpus size is more than sufficient for performing machine learning studies.

2. Factors used in modeling referential choice: The full set of features

We are using the following set of factors of referential choice. Most of these factors were already mentioned in (Kibrik et al. 2010); we italicize below those factors that were added later on. Where appropriate, we indicate in parentheses the technical terms used for the factors in the annotation scheme.

Referent's features:

- Animacy: animate (human) or inanimate (non-human)
- *Gender and number* (agreement)*: masculine, feminine, neuter, plural*
- Protagonism, that is a referent's centrality in discourse (see below)

Antecedent's features:

- Affiliation in direct speech (dir_speech); this feature is relevant both for the anaphor and the antecedent, because particularly important are the situations in which they are located across a direct speech boundary
- Type of phrase (phrase_type): noun phrase, prepositional phrase, other
- Grammatical role (gramm_role): subject, direct object, indirect object, other
- Referential form (np_form, def_np_form): definite NP, with further indication of subtype, vs. proper name vs. indefinite NPs
- *Antecedent length, in words*
- *Number of markables from the anaphor back to the nearest full NP antecedent*

Anaphor's features:

- Introductory vs. repeated mention (referentiality)
- *Number of referent mention in the referential chain*
- Affiliation in direct speech (dir_speech)
- Type of phrase (phrase_type): noun phrase, prepositional phrase, other
- Grammatical role (gramm_role): subject, direct object, indirect object, other

Distances between anaphor and antecedent:

- Distance in words
- Distance in markables; this feature partly accounts for referential competition in a discourse context, that is issues related to potential ambiguity or referential conflict (see Kibrik 1987)
- Linear distance in elementary discourse units, as found in the rhetorical representation
- Rhetorical distance in elementary discourse units, as found in the rhetorical representation
- *Distance in sentences*
- *Distance in paragraphs*.

Recently we have given particular attention to modeling the factor of referent's protagonism in discourse. For this goal referential chains were identified, that is sequences of referential expressions naming the same referent. Each

referential chain has a certain length, that is, the number of referential expressions it contains.

Two models of protagonism were used. In the first one, to each referent corresponds the ratio of its referential chain length to the maximal length of a referential chain in the text. In the second model, to each referent corresponds the ratio of its referential chain to the gross number of markables in the text. In both instances the most frequently mentioned referent is the same, but relative weights of referents may be different.

In order to test to which extent a given model of protagonism corresponds to human text understanding, experiments were undertaken (Linnik 2010). Thirty texts were chosen from the RST Discourse Treebank. The length of the texts varied from 70 to 1344 words. Experiment participants were required to read the text and to identify the central entity (protagonist). Each text was analyzed by three experiment participants.

Experiment participants were thirty native speakers of English, from 20 to 54 years of age. For 50% of the texts, namely 15, all participants were unanimous in choosing the protagonist. Eleven more texts showed the agreement between two (out of three) participants in their protagonist assessment. That is, 26 texts out of 30 (87%) provide relatively reliable information on human-selected protagonists.

A comparison of the experiment results with the results of computational analysis demonstrated that the human assessment and the computer's assessment coincide in 24 instances out of 26. Therefore, the automatic models predict human identification of protagonist 92% of the time.

One more factor that deserves special mention is the factor of rhetorical distance. There are several complications in how this measurement is applied to various rhetorical configurations. These complications were discussed in (Kibrik, Krasavina 2005); in the current project we followed the methods proposed in that study.

3. Interaction between factors: methods of computer learning

In the computational model of referential choice the following two tasks were set:

- to predict whether a given anaphor is a (third person) pronoun or a full noun phrase (two-way task)
- to predict whether a given anaphor is a (third person) pronoun or a descriptive noun phrase or a proper name (three-way task).

From the beginning of this project, several algorithms of machine learning were chosen, belonging to different groups: logical classifiers and logistic regression (Kibrik et al. 2010). The results of the logical algorithms (decision trees C4.5, deciding rules algorithm JRip) lend themselves to natural interpretation. Logistic regression was chosen for the following two reasons. First, the results of this algorithm excel those of logical algorithms in quality. Second, logistic regression allows one to obtain probabilistic estimates of referential options.

More recently, we also used the so-called classifier compositions: bagging and boosting.

The boosting algorithm (Freund, Schapire 1996) uses as its parameter another machine learning algorithm that we will call the base algorithm. The base algorithm undergoes optimization. An adaptation of classifiers is performed, that is, each additional classifier applies to the objects that were not properly classified by the already constructed composition. After each call of the algorithm the distribution of weights is updated. (These are weights corresponding to the importance of the training set objects.) At each iteration the weights of each wrongly classified object increase, so that the new classifier focuses on such objects. Among the boosting algorithms, AdaBoost was used in our modeling with the C4.5 base algorithm.

Bagging (from "bootstrap aggregating"; Breiman 1994) algorithms are also algorithms of composition construction. Whereas in boosting each algorithm is trained on one and the same sample with different object weights, bagging randomly selects a subset of the training samples in order to train the base algorithm. So we get a set of algorithms built on different, even though potentially intersecting, training subsamples. A decision on classification is done through a voting procedure in which all the constructed classifiers take part. In the case of bagging the base algorithm was also C4.5.

In the current set of modeling studies we used 4291 anaphor-antecedent pairs, including 2854 full noun phrases and 1437 pronouns as anaphors. In order to control the quality of classification, the cross-validation procedure was used:

1. The training set is divided into 10 parts.
2. A classifier operates on the basis of 9 parts.
3. The constructed decision function is tested on the remaining part.

The procedure is repeated for all possible partitions, and the results are subsequently averaged. The criterion for choosing both an optimal set of features and an algorithm is **accuracy**, that is the ratio of properly predicted referential expressions to the overall amount of referential expressions.

The results of modeling studies are given in Table 1 (two-way task) and Table 2 (three-way task). In the columns "Accuracy 2010" results are provided for the set of factors included in (Kibrik et al. 2010), whereas the columns "Accuracy 2011" include the new factors incorporated into the model at the more recent stage.

Table 1. Modeling referential choice in the two-way task: full noun phrase vs. pronoun

Algorithm	Accuracy 2010	Accuracy 2011
Logistic regression	85.6%	87.0%
Decision tree algorithm	84.3%	86.3%
Deciding rules algorithm	84.5%	86.2%
Boosting	88.2%	**89.9%**
Bagging	86.6%	87.6%

Table 2. Modeling referential choice in the three-way task: descriptive noun phrase vs. proper name vs. pronoun

Algorithm	Accuracy 2010	Accuracy 2011
Logistic regression	76.0%	77.4%
Decision tree algorithm	74.3%	76.7%
Deciding rules algorithm	72.5%	75.4%
Boosting	79.3%	80.7% — 50 iterations **80.9%** — 100 iterations
Bagging	78.0%	79.5% — 50 iterations 79.6% — 100 iterations

Thus the enlistment of new features in the recent modeling studies, as well as the use of additional algorithms of machine learning, allowed us to noticeably improve the prediction of referential choice.

4. Significance of factors and factor correlations

As was shown in section 2, six different distance measurements were used. In order to find out which of the distances correlate with each other, the Spearman's correlation coefficient was computed that reveals linear dependencies between variables. If two variables have the Spearman's coefficient of 1, they are in a linear dependency. If the coefficient value is −1, there is an inverse dependence. The coefficient values obtained for all pairs of distances are shown in Table 3.

Table 3. Correlations between different anaphor–antecedent distances

Distance in:	Words	Markables	Elementary discourse units (linear)	Elementary discourse units (rhetorical)	Sentences	Paragraphs
Paragraphs	0.6629	0.5617	0.6538	0.6169	0.7734	1.0000
Sentences	0.7663	0.6034	0.7530	0.6569	1.0000	
Elementary discourse units (rhetorical)	0.5864	0.4746	0.6598	1.0000		
Elementary discourse units (linear)	0.8748	0.6753	1.0000			
Markables	0.7051	1.0000				
Words	1.0000					

As can be seen from Table 3, the maximal correlation is observed between the distance in words and the linear distance in elementary discourse units, while the minimal correlation is observed between rhetorical distance and the distance in words. Minimally correlated with other types of distance are rhetorical distance and the distance in markables. Note, however, that the cognitive interpretation of the distance in markables is yet to be determined.

Also, for the three-way task the results of classification were computed with the deduction of certain factors and groups of factors, see Table 4. An analysis of the contribution of newly added factors was also performed.

Table 4. The significance of factors in the three-way task of referential choice

Factors	Accuracy
All factors, including the newly added ones (boosting with 50 iterations)	**80.7%**
without protagonism	80.0%
without affiliation in direct speech, for both anaphor and antecedent	80.6%
without animacy	80.68%
without all distances	73.5%
— except for the distance in words only	79.0%
— except for rhetorical distance only	74.9%
— except for the distances in words and paragraphs	79.0%
— except for the distances in words and sentences	79.5%
— except for the distances in words, sentences, and paragraphs	79.4%
— except for rhetorical distance and the distances in words and sentences	79.7%
— except for the rhetorical distance and the distances in words and markables	79.9%
— except for the distances in words, markables, and paragraphs	80.47%
without the anaphor's grammatical role	79.3%
without the antecedent's grammatical role	80.2%
without grammatical role	79.2%
without the antecedent's referential form	77.0%
Old factors (Kibrik et al. 2010) (boosting with 50 iterations)	**79.3%**
plus referent number and gender	79.7%
plus number of markables to the nearest full NP plus chain length	78.9%
plus antecedent length	78.7%
plus distance in sentences	79.5%
plus distance in paragraphs	79.25%
plus antecedent gender plus distance in paragraphs plus distance in sentences	80.3%

Table 4 makes explicit the significance of various factors, such as different distance measurements, protagonism, grammatical role, antecedent's referential form, etc. Note that the inclusion of the distance in markables leads to the improvement of classification (underscored in Table 4). Perhaps this is due to the fact that this factor indeed helps to take into account referent competition or referential conflict.

The analysis of the data in Table 4 demonstrates that the great majority of the factors are significant and cannot be easily removed from the model. Even the numerous distance measurements do not lend themselves to substantial reduction.

Conclusion

In this paper we have presented the recent results of our modeling study in referential choice, based on the RefRhet corpus. The account of additional factors and the employment of compositions of machine learning techniques have led to an improvement of referential choice prediction. This applies both to the two-way choice between full NP and pronoun and to the three-way choice "descriptive NP vs. proper name vs. pronoun". We have demonstrated that the great majority of the factors taken into account are significant for modeling referential choice.

References

1. *Belz A., Kow E., Viethen J., Gatt A.* 2008. The GREC Challenge: Overview and Evaluation Results. Proceedings of the Fifth International Natural Language Generation Conference : 183–191.
2. *Breiman L.* 1994. Bagging Predictors Technical Report 421, Department of Statistics.
3. *Byron D. K., Gegg-Harrison W.* 2004. Eliminating Non-referring Noun Phrases from Coreference Resolution. Proceedings of the Discourse Anaphora and Anaphora Resolution Confenrence (DAARC2004) : 21–26.
4. *Carlson L. D., Marcu D., Okurowski M. E.* 2003. Building a Discourse-tagged Corpus in the Framework of Rhetorical Structure Theory. Current Directions in Discourse and Dialogue : 85–112.
5. *Fox B.* 1987. Discourse Structure and Anaphora in Written and Conversational English.
6. *Freund Y., Schapire R.* 1996. Experiments with a New Boosting Algorithm. Mashine Learning: Proceedings of the Thirteenth International Conference.
7. *Ge N., Hale J., Charniak E.* 1998. A Statistical Approach to Anaphora Resolution. Proceedings of the Sixth Workshop on Very Large Corpora : 161–170.
8. *Givón T.* 1983. Topic Continuity in Discourse: An Introduction. Topic Continuity in Discourse: A Quantitative Cross-language Study : 1–42.
9. *Grüning A., Kibrik A. A.* 2005. Modeling Referential Choice in Discourse: A Cognitive Calculative Approach and a Neural Networks approach. Anaphora Processing: Linguistic, Cognitive and Computational Modelling : 163–198.

10. *Kibrik A. A.* 1987. Mechanisms of Referential Conflict Removal [Mekhanizmy Ustraneniia Referentsial´nogo Konflikta]. Modelirovanie Iazykovoi Deiatel'nosti v Intellektual'nykh Sistemakh :128–145.
11. *Kibrik A. A.* 1996. Anaphora in Russian Narrative Discourse: A Cognitive Calculative Account. Studies in Anaphora :255–304.
12. *Kibrik A. A.* 1999. Reference and Working Memory: Cognitive Inferences From Discourse Observation. Discourse Studies in Cognitive Linguistics : 29–52.
13. *Kibrik A. A., Dobrov G. B., Zalmanov D. A., Linnik A. S., Loukachevitch N. V.* 2010. Referential Choice as a Multi-factor Probabilistic Process [Referencial'nyi Vybor kak Mnogofaktornyi Veroiatnostnyi Protsess]. Computational Linguistics and Intellectual Technologies. Papers from the Annual International : 173–181.
14. *Krasavina O., Chiarcos Ch.* 2007. PoCoS — Potsdam Coreference Scheme. Proceedings of the Linguistic Annotation Workshop (LAW) :156–163.
15. *Linnik A. S.* 2010. Linguistic Support for Computational Analysis of the Corpus of Texts, Annotated with Respect to the Referential Theory.
16. *Mann W. C., Thompson, S. A.* 1988. Rhetorical Structure Theory: Toward a functional theory of text organization, 8(3): 243–281.
17. *Nenkova A.* 2008. Entity-driven Rewrite for Multi-Document Summarization. Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP) : 118–125.
18. *Orasan C.* 2004. The Influence of Personal Pronouns for Automatic Summarization of Scientific Articles. Proceedings of the Discourse Anaphora and Anaphora Resolution Confenrence (DAARC2004) :127–132.
19. *Strube M., Wolters M.* 2000. A Probabilistic Genre-independent Model of Pronominalization. Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics : 18–25.
20. *Tetreault J. R.* A Corpus-Based Evaluation of Centering and Pronoun Resolution. Computational Linguistics, 27(4) : 507–520.