Computational Modeling Of Referential Choice: Major And Minor Referential Options

Mariya V. Khudyakova (mariya.kh@gmail.com)
Moscow State University, Faculty of Philology,
Department of Theoretical and Applied Linguistics,
GSP-1, Leninskie Gory, Moscow, 119991 Russia

Grigory B. Dobrov (wslc@rambler.ru)
Moscow State University, Faculty of Computational Mathematics and Cybernetics,
GSP-1, Leninskie Gory, Moscow, 119991 Russia

Andrej A. Kibrik (aakibrik@gmail.com)
Institute of Linguistics, Russian Academy of Sciences,
1-1 B.Kislovskiy per., Moscow, 125009, Russia

Natalia V. Loukachevitch (louk@mail.cir.ru)
Moscow State University, Faculty of Computational Mathematics and Cybernetics,
GSP-1, Leninskie Gory, Moscow, 119991 Russia

Abstract

The choice between different types of referential expressions, such as definite descriptions, proper names and pronouns, depends on a large number of factors acting simultaneously. In this study the role and the significance of these factors is modeled with the help of different algorithms of machine learning. The work is based on the specially designed RefRhet corpus with anaphora annotation. The paper focuses on predicting major referential options (such as the-NPs and pronouns) and minor options (such as demonstrative NPs).

Keywords: referential choice, corpus annotation, modeling, machine learning

Introduction

Performing reference, that is, naming persons or objects, is the necessity every speaker or writer faces constantly when producing discourse. Speakers and writers need to choose an appropriate referential expression from a repertoire of several major forms of reference, including pronouns, descriptive noun phrases, and proper names. This procedure is called referential choice.

The project reported in this paper relies on a number of cognitively-oriented studies of referential choice, such as (Chafe, 1994; Fox, 1987; Givón, 1983), and the assumption that the choice of a referential expression depends on the status of the referent in the speaker's cognititive system (working memory). More specifically, this study is based on the multi-factorial approach to referential choice developed by Andrej A. Kibrik in a series of publications; see (Kibrik, 2011) for the most up-to-date version. In the earlier publications Kibrik (1999, 1996) proposed a calculative model of referential choice that included a number of factors, each having a set of values with certain numerical weights. The sum of the weights was supposed to measure the activation of the referent in the speaker's working memory and predict referential choice between a full NP and a pronoun. Grüning and Kibrik (2005) attempted a model of referential choice in which the weights of individual factors were 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.

These studies were based on relatively small data sets. A much larger corpus is needed in order to implement a fully-fledged machine learning-based study. Large corpora have been used before in reference studies. For example, one corpus of this kind was formed for the GREC conference, see e.g. http://www.nltg.brighton.ac.uk/research/genchal10/grec/.

RefRhet corpus

In the studies (Loukachevitch, Dobrov, Kibrik, Khudyakova & Linnik, 2011; Kibrik, Dobrov, Zalmanov,

Linnik & Loukachevitch, 2010) we described a computational model of referential choice, based on the specially designed RefRhet corpus.

In this paper we present new results of our project and also address the problem of predicting demonstrative full NPs.

The RefRhet corpus is based on the English-language corpus RST Discourse Treebank (http://www.isi.edu/~marcu/discourse/Corpora.html), see (Carlson, Marcu & Okurowski, 2003). This corpus contains annotation for rhetorical structure). As was demonstrated by Fox (1987), rhetorical structure is important for reference in discourse; on the basis of this idea Kibrik (1996) proposed the measurement of rhetorical distance that proved significant in the studies of referential choice. RST Discourse Treebank consists of 385 Wall Street Journal articles and contains 176 383 words.

Referential annotation was added to the RST Discourse Treebank, and as a result the RefRhet corpus was formed. Text constituents that serve as referential expressions (markables) have been annotated in the corpus. Coreference relations are posited between markables. Any non-first mention *n* of a referent (that is, anaphor) is connected with the previous mention *n–1* (that is, antecedent). In addition, each markable contains a number of annotated features that can affect referential choice. The present-day annotation is done with the help of an annotation scheme (Krasavina & Chiarcos, 2007) using the MMAX-2 program, created specifically for modeling reference (see http://mmax2.sourceforge.net/).

The procedure requires the annotation of each text to be performed twice by two different annotators, in order to minimize errors. Then the two variants are compared automatically. Such comparison results in a list of markables that either appear in one of the annotations only, 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 size of the RefRhet corpus is comparable with the other corpora of this kind that exist to date; cf. (Rodriguez, Delogu, Versley, Stemle, & Poesio, 2010; Poesio & Artstein, 2008); GREC corpora. Given that the annotation of a referential corpus is an extremely laborious task, creating a larger corpus would simply be unpractical. From a statistical point of view, the corpus size is more than sufficient for performing machine learning studies.

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

We currently use the set of 20 factors, including the referent's features, antecedent's and anaphor's features and the distances between the anaphor and the antecedent.

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

Antecedent's features:

- Affiliation in direct speech; this feature is relevant both for the anaphor and the antecedent; particularly important are the situations in which the anaphor and the antecedent appear across a direct speech boundary
- Type of phrase: noun phrase, prepositional phrase, other
- Grammatical role: subject, direct object, indirect object, other
- Referential form: definite NP, with further indication of subtype, proper name, indefinite NPs, pronouns, with further identification of subtype
- Antecedent length, in words
- Number of markables from the anaphor back to the nearest full NP antecedent

Anaphor's features:

- Introductory vs. repeated mention
- Number of referent mentions in the referential chain
- Affiliation in direct speech
- Type of phrase: noun phrase, prepositional phrase, other
- Grammatical 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, 2011)
- Linear distance in elementary discourse units, roughly equaling clauses
- Rhetorical distance in elementary discourse units, as found in the rhetorical representation

- Distance in sentences
- Distance in paragraphs.

The analysis of the referring expressions prediction in (Kibrik et al. 2011) demonstrated 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.

Predicting the Choice between the major referential options

In the computational model of referential choice the following two tasks were initially 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 full noun phrase or a proper name (three-way task).

We use several algorithms of machine learning: decision trees C4.5, deciding rules algorithm JRip, and logistic regression (see Hall, Eibe, Holmes, Pfahringer, Reutemann, & Witten, 2009), as well as the so-called classifier compositions: bagging (Breiman, 1994) and boosting (Freund & Schapire, 1996).

In the present study we use a subcorpus containing 4291 anaphor-antecedent pairs, including 2854 full noun phrases and 1437 pronouns as anaphors.

The results of modeling for the two-way task and the three-way task are given in Table 1 and Table 2.

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

Algorithm	**Accuracy**
Logistic regression	87.0%
Decision tree algorithm	86.3%
Deciding rules algorithm	86.2%
Boosting	**89.9%**
Bagging	79.6%

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

Algorithm	**Accuracy**
Logistic regression	77.4%
Decision tree algorithm	76.7%
Deciding rules algorithm	75.4%
Boosting	**80.9%**
Bagging	87.6%

Predicting minor referential options

In addition to the coarse choice between the major referential options that include pronouns and two sorts of full NPs, we address the use of minor referential options that occur with a significantly lower frequency in natural discourse. Specifically, we focus on demonstrative full NPs, such as *this company* or *that company*, as opposed to plain descriptive full NPs with the definite article (the-NPs for short), such as *the company*.

The choice of definite descriptions subtypes: the-NPs and demonstrative NPs

When predicting the choice of definite description subtypes we assume that the choice of a full NP has already been made, and the algorithms model only the choice between the-NPs and demonstrative NPs (we call here the choice options "classes").

The task of predicting the subtype of definite description is in a way more complicated than modeling the choice between major referential options. This is because the frequency of demonstrative NPs is very low compared to the major referential options. For example, the subcorpus contains 1083 the-NPs and 72 demonstrative NPs with antecedents. So demonstrative NPs make only 6% of all definite descriptions. As can be seen from Table 3, the choice between the two kinds of descriptive NPs is predicted with high accuracy, but this is simply due to the fact that the algorithms always predict the dominant option (the-NP). Only from 0% to 31% of demonstrative full NPs were correctly predicted by the algorithm.

Table 3: Modeling referential choice between the-NPs and demonstrative NPs.

Algorithm	Accuracy		
	Total	The-NPs	Demonstrative full NPs
Decision tree algorithm	94%	100%	0%
Logistic regression	94%	98%	31%
Boosting	94%	99%	19%

The errors are distributed unevenly between the two classes. We have used the Metacost method to solve this problem. Metacost allows to set different penalties on errors in different classes, and then use the given algorithms (Domingos 1999).

The results of predicting the subtype of definite NPs using Metacost with the penalty coefficients 1:10 and 1:15 are given in Table 4 and Table 5.

These results show that, with the use of Metacost, we have gained a relatively high quality of predicting the subtypes of definite descriptions.

Table 4: Modeling referential choice in the two-way task: the-NPs vs. demonstrative NPs using Metacost (1:10 penalty coefficients)

Algorithm	Accuracy	
	The-NPs	Demonstrative full NPs
Decision tree algorithm	95%	25%
Logistic regression	**81%**	**57%**
Boosting	99%	22%

Table 5: Modeling referential choice in the two-way task: the-NPs vs. demonstrative NPs using Metacost (1:10 penalty coefficients)

Algorithm	Accuracy	
	The-NPs	Demonstrative full NPs
Decision tree algorithm	86%	44%
Logistic regression	**74%**	**68%**
Boosting	99%	22%

Three-way task: the-NPs, third person pronouns and demonstrative NPs.

Since using Metacost proved to be helpful in predicting minor referential options, we have used it to model the three-way choice between the-NPs, third person pronouns and demonstrative NPs.

Table 6: Modeling referential choice between the-NPs, third person pronouns and demonstrative NPs using Metacost (1:15 penalty coefficients).

Algorithm	Accuracy			
	Total	The-NPs	Third person pronouns	Demonstrative NPs
Decision tree algorithm	83%	79%	88%	24%
Logistic regression	**80%**	**73%**	**90%**	**57%**
Boosting	90%	90%	93%	21%

This is a very complicated task since the numbers of the-NPs and pronouns are high and rather close (1083 and 1136 respectively), whereas the number of demonstrative NPs is very low (72). Table 6 shows the results of predicting the type of referential expressions (three-way task) using Metacost with the penalty coefficients 1:15. The results of the Metacost-based prediction with the penalty coefficients 1:10 and 1:20 are somewhat lower.

Conclusions

We have reached substantial success in modeling the coarse choice between the major, most frequent referential options as a multifactorial process. A relatively large corpus, containing annotation for many potentially relevant factors has been created, and set of machine learning techniques were trained to predict the choice between descriptive NPs, proper names, and pronouns. The accuracy of prediction for the two-way and the three-way tasks is in the vicinity of 90% and 80%, respectively.

The prediction of minor referential options requires the utilization of special machine learning techniques, such as Metacost. Using metacost increases the quality of prediction of referential choice when predicting the subtype of definite descriptions as well as in the task involving a minor referential option along with major referential options.

In future research we intend to further refine the machine learning methods, to bring into picture other minor referential options and thus come up with a more complete model of referential choice.

Aknowledgements

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

References

Breiman, L. (1994). *Bagging Predictors* (Technical Report 421), Department of Statistics, University of California at Berkeley.

Carlson L.D., Marcu D. and Okurowski M.E. (2003). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith (eds.) *Current directions in discourse and dialogue* (pp. 85–112). Dordrecht: Kluwer, 2003.

Chafe, W.L. (1994). *Discourse, consciousness and time.* Chicago: University of Chicago Press.

Domingos P. (1999) Metacost: A general method to make classifiers cost-sensitive. *Proceedings of the Fifth*

International Conference on Knowledge Discovery and Data Mining (pp. 155-164)

Fox B. (1987). *Discourse structure and anaphora in written and conversational English*. Cambridge: Cambridge University Press

Freund, Y. & Schapire, R. (1996). Experiments with a New Boosting Algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*.

Givón T. (1983) Topic continuity in discourse: An introduction. In T. Givón (Ed.), *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam: Benjamins

Grüning, A. & Kibrik, A. A. (2005). Modeling referential choice in discourse: A cognitive calculative approach and a Neural Networks approach. In A. Branco, T. McEnery and R. Mitkov (eds.). *Anaphora processing: Linguistic, cognitive and computational modelling*. Amsterdam: Benjamins, 2005. Pp. 163–198.

Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, Ian H.. (2009). The WEKA data mining software: An update. *SIGKDD Explorations, Volume 11, Issue 1*.

Kibrik, A. A. (1996). Anaphora in Russian narrative discourse: A cognitive calculative account. In B. Fox (ed.) *Studies in anaphora*. Amsterdam: Benjamins, 1996.

Kibrik, A. A. (1999). Reference and working memory: Cognitive inferences from discourse observation. In K. van Hoek, A. A. Kibrik & L. Noordman (Eds.) *Discourse studies in cognitive linguistics*. Amsterdam: Benjamins.

Kibrik A.A. (2011). *Reference in discourse*. Oxford: Oxford University Press, 2011, in press.

Kibrik A.A., Dobrov G.B., Zalmanov D.A., Linnik A.S. and Loukachevitch N.V.. (2010). Referencial'nyj vybor kak mnogofaktornyj verojatnostnyj process [Referential choice as a multi-factor probabilistic process]. In A.E. Kibrik (ed.), *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference Dialogue 2010*. Bekasovo, Moscow region. Moscow: RGGU.

Krasavina, O., Chiarcos, Ch. (2007). PoCoS — Potsdam Coreference Scheme. In *Proceedings of the Linguistic Annotation Workshop (LAW)* (pp. 156–163). June 28–29, 2007, Prague, Czech Republic. Stroudsburg, PA: Association for Computational Linguistics

Loukachevitch N. V., Dobrov G.B., Kibrik A.A., KhudyakovaM.V. &Linnik A.S. (2011). Factors of referential choice: computational modeling. In A.E. Kibrik (ed.), *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference Dialogue 2011*. Bekasovo, Moscow region. Moscow: RGGU.

Poesio, M. & Artstein, R. (2008). *Anaphoric annotation in the ARRAU corpus*. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008.

Rodriguez, K. J., Delogu, F., Versley, Y., Stemle, E. and Poesio, M.. (2010). Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May 2010.