A NEURAL NETWORK APPROACH TO REFERENTIAL CHOICE

André Grüning
Max-Planck-Institute for Mathematics in the Sciences, Leipzig
Inselstr. 22—26, 04105 Leipzig, Germany
gruening@mis.mpg.de
Andrej A. Kibrik
Institute of Linguistics, Russian Academy of Sciences
B. Kislovskij per. 1/12, Moscow 103009 Russia
kibrik@comtv.ru

Abstract

In Kibrik's (1996, 1999, 2000) sample-based studies of referential choice, a quantitative approach was proposed to calculate an activation score of a specific referent at a particular point of discourse from a range of partly interdependent factors, such as distance to the antecedent, referent animacy, etc. The activation score then determines the referential choice. The advantage of that multi-factorial approach is that it is explanatory and testable. That approach, however, was mathematically rather unversed and had some shortcomings. The list of factors and their interaction to yield the activation score were hand-coded. The interaction was purely additive, ignoring possible non-linear interdependencies between the factors. In this paper we propose a more sophisticated neural network analysis of the same data. We trained a feed-forward network on the data. It classified 96% of all data correctly with respect to the actual referential choice. A pruning procedure allowed to produce a minimal network and revealed that out of ten input factors five were sufficient to predict the data almost correctly, and that the logical structure of the remaining factors can be simplified. This is a pilot study necessary for the preparation of a larger neural network-based study. The neural network approach allows to model the set of necessary and sufficient activation factors explaining referential choice in discourse. Its big advantage over classical statistical methods is that the type of regularities it can detect in the data is less constrained.

1. Introduction

We approach the phenomena of discourse reference as a realization of the process of *referential choice*: every time the speaker needs to mention a referent s/he has a variety of options at his/her disposal, such as full NPs, demonstratives, third person pronouns, etc. The speaker *chooses* one of these options according to certain rules that are a part of the language production system. Production-oriented accounts of reference are rarer in the literature than comprehension-oriented; for some examples see Dale (1992), Strube and Wolters (2000).

Linguistic studies of referential choice often suffer from circularity: for example, a pronominal usage is explained by the referent's high activation, while the referent is assumed to be highly activated because it is actually coded by a pronoun in discourse. In a series of studies by Kibrik (1996, 1999, 2000) an attempt to break that circularity was undertaken. The main methodological idea is that we need an account of referent activation that is entirely independent of the actual referential choices observed in actual discourse. There are a variety of linguistic factors that determine a referent's current activation, and once the level of activation is determined, the referential option(s) can be predicted with a high degree of certainty.

The approach proposed in the aforementioned studies includes a quantitative component that models the interaction of activation factors yielding the summary activation of a referent. As will be explained below, the contributions of individual factors are simply summed, and for this reason we will use the shorthand *calculative approach*, although the quantitative component is only one facet of that approach.

In this pilot study we intend to find out whether neural networks can help us to overcome some shortcomings of the calculative approach. As the available data set is quite small (102 items) and large annotated corpora are not so easily obtained, we decided to design this study as a case study, rather than putting weight on statistical rigor.

2. The calculative approach and its properties

The approach proposed by Kibrik (1996, 1999, 2000) needs to be briefly outlined. That approach is:

- *speaker-oriented*: referential choice is viewed as a part of language production performed by the speaker
- *sample-based*: the data for the study is a sample of natural discourse, rather than heterogeneous examples from different sources
- *general*: all referential devices in sample must be accounted for
- *finite*: the proposed list of factors cannot be supplemented to account for exceptions
- *predictive*: the proposed list of factors is supposed to predict referential choice with 100% certainty
- *explanatory and cognitively based*: it is claimed that this approach models the actual cognitive processes, rather than relies on a black box ideology
- *multi-factorial*: multiplicity of factors determining referential choice is recognized
- *calculative*: contributions of activation factors are numerically characterized
- *testable*: all components of this approach are subject to verification

The calculative approach has the following structure. The primary cognitive determiner of referential choice is *activation* of the referent in question in the speaker's working memory, as well as in the addressee's working memory, as assessed by the speaker (cf. Chafe, 1994, Tomlin and Pu, 1991). At any given moment any referent has a certain level of activation, so-called *activation score*. Activation score depends on a number of factors rooted in discourse context (for example, distance to the antecedent) and in the referent's properties (for example, protagonisthood in the current discourse). 7 and 11 activation factors have been identified for Russian (Kibrik, 1996) and English (Kibrik, 1999), respectively. Factors differ in their numeric contribution very much. For example, rhetorical distance to the antecedent (first introduced in Fox, 1987) can contribute up to 0.7 to activation score, while protagonisthood can contribute maximally 0.2. Given that the value of each activation factor can be identified for each referent at any time, numeric weights of activation factors are added and give rise to activation score. Numeric weights of activation factors have been identified by hand, through a complicated trial-and-error procedure. Three selected activation factors, with their values and numeric weights, are shown in Table 1.

Table 1: Activation factors and their values, as identified for English narrative discourse

Activation factors	Values of activation factors	Corresponding numeric weights
Rhetorical distance to the antecedent (RhD)	1	0.7
	2	0.5
	3+	0
Paragraph distance to the antecedent (ParaD)	0	0
	1	–0.3
	2+	–0.5
Protagonisthood	No	0
	Yes, and	
	RhD+ParaD ≤ 2	0
	RhD+ParaD ≥ 3	0.2

Referent's activation score varies within a certain range (e.g. between 0 and 1). If the current activation score is above a certain *threshold*, then a semantically reduced (pronoun or zero) reference is possible, and if not, a full NP is used. In Russian and English, by far the most important referential devices are third person pronouns and definite full NPs. There is an important difference between alterable and categorical referential devices. For examples, some third person pronouns can vary with full NPs (alterable), and some others cannot (categorical). Different activation levels correspond to such different types of pronouns.

Mappings from activation levels onto referential devices are called *referential strategies*. The English referential strategies identified in Kibrik (1999) are shown in Table 2. (The quantitative system is set up so that the activation score can vary between 0 and 1.2.)

Table 2: Referential strategies in English narrative discourse

Referential device:	Full NP only	Full NP, ?pronoun	Either full NP or pronoun	Pronoun, ?full NP	Pronoun only
Activation score:	0–0.2	0.3–0.5	0.6–0.7	0.8–1.0	1.1+

3. Shortcomings of the calculative approach

There are some problems with the calculative approach, especially with its quantitative component per se that was mathematically quite unversed.

First, the list of relevant activation factors may not be necessary and sufficient. Those factors were included in the list that showed a strong correlation with referential choice. However, only all factors in conjunction determine the activation score, and therefore the strength of correlation of individual factors may be misleading, and the contribution of individual factors is not so easy to identify.

Second, numeric weights of individual factors' values were chosen by hand, and again are not necessarily optimal. The weights of activation factors were intrinsically subjective in that approach.

Third, the interaction between factors was purely additive, ignoring possible non-linear interdependencies between the factors. Non-linear dependencies are particularly probable, given that some factors interact with others, such as protagonisthood interacts with rhetorical and paragraph distance (see Table 1). Other factors might be correlated, e.g. animacy, protagonisthood, and the syntactic role of subject. Also, from the cognitive point of view it is unlikely that such a simple procedure as mere addition can adequately describe processing of activation in the brain.

Fourth, because of the additive character of factor interaction it was very hard to limit possible activation by a certain range. It would be intuitively natural to posit that minimal activation varies between zero and some maximum which can without loss of generality be assumed to be one. However, because of penalizing factors such as paragraph distance that deduct activation it often happens that activation score turns out negative, which makes a cognitive interpretation difficult.

In order to solve these problems, the idea to develop a more sophisticated mathematical apparatus emerged, such that:

- identification of significant factors, numeric weights, and factor interaction would all be interconnected and would be a part of the same task
- the modeling of factors would be done computationally, by building an optimal model of factors and their interaction.

4. Proposed solution: a neural network approach

In the neural network approach, we lift the requirement of complete predictiveness: we posit that referential choice can predict/explain referential choice with a degree of certainty that can be less than 100%.[1] Also, the neural network approach is different from the previous approach in that it does not make specific claims about cognitive adequacy and activation. There is no such thing as summary activation score in this approach at its present stage. Activation factors themselves are reinterpreted as mere parameters or variables in the data that are mapped onto referential choice. Perhaps at a later stage the neural network approach can embrace the explanatory cognitive component. Its big advantage over classical statistical methods is that the type of regularities it can detect in the data is less constrained.

The term *artificial neural network* or *net* denotes a variety of different function approximators that are neuro-biologically inspired (Fine, 1999). Their common property is that they can, in a supervised or unsupervised way, learn to classify data. A net consists of *nodes* that are connected by *weights*. Every node integrates the activation it gets from its predecessor nodes in a non-linear way and sends it to its successors. The nodes are ordered in layers. Numerical data is presented to the nodes in the *input layer*, from where the activation is injected into one or more hidden layers, where the actual computation is done. From there activation spreads to the *output layer,* where the result of the computation is read off.

For this pilot study we decided to employ a simple feed-forward network with the back-propagation learning algorithm. In this supervised learning task the network must learn to predict from ten factors (Table 3), whether the given referent will be realized as a pronoun or a full noun phrase. To input the factors with symbolic values into the net, they have to be converted into numeric values. If the symbolic values denote some gradual property such as animacy, they are converted into one real variable with values between –1 and 1. The same holds true for binary variables. If there was no a priori obvious order in

[1] But this might also be a desirable feature, e.g. to account for stylistic variation.

the symbolic values, they were coded unary (e.g. Syntactic Role), i.e. to every value of that factor corresponds one input node, which is set to one if the factor has this value and to zero otherwise.

Thus 24 input nodes and one output node are needed. The output node is trained to predict, whether the actual referent under consideration is realized as a full noun phrase (FNP) (numerical output below 0.4) or as a pronoun (pro) (numerical output above 0.6).[2]

Table 3. Factors used in Simulation 1, their possible values and the corresponding input nodes.

S, Poss, Obl, DO, IOag mean Subject, Possessor, Oblique, Direct object and Agentive indirect object. Pred means Predicative use, Pro – pronoun and FNP – full noun phrase

Factor	Values	Coding	Input Nodes
Syntactic Role	S, Poss, Obl, DO, IOag	Unary	1—5
Animacy	Human, animal, inanimate	Human: 1, animal: 0, inanimate: –1	6
Protagonisthood	Yes / no	Binary	7
Syntactic Role of Rhetorical Antecedent	S, Poss, Obl, Pred, DO, IOag	Unary	8—13
Type of Rhetorical Antecedent	Pro, FNP	Binary	14
Syntactic Role of Linear Antecedent	S, Poss, Obl, Pred, DO, IOag	Unary	15-20
Type of Linear Antecedent	Pro, FNP	Binary	21
Linear Distance to Antecedent	Integer	Integer	22
Rhetorical Distance to Antecedent	Integer	Integer	23
Paragraph Distance to Antecedent	Integer	Integer	24

All – at this point – numerical input values were normalized to have zero mean and unit variance. This normalization was done to ensure that all data a priori is treated on equal footing and the impact of a factor can be directly read off from the strength of the weight connecting its input node to the hidden or output layer.

4.1. Simulation 1 – full data set

A network with 24 nodes in a single hidden layer was trained on the data set of 102 items from Kibrik (1999) for 1000 epochs.[3] As parts of the training are stochastic that experiment was repeated several times. In all cases the net learned to predict the data correctly except for a small number (below six) cases. Typically, the misclassifications occurred for the same items in the data set. A closer analysis of a well trained net with only four misclassifications revealed that three of them are due to referential conflict (which was not among the input factors), that is, in the situation when the full noun phrase is used only because a pronoun (otherwise expected) may turn out ambiguous.

4.2. Simulation 2 – pruning

Not only did we want our net to learn the data but also to make some statements about the importance of the input factors and their interdependency. To achieve this goal we subjected the trained net from Simulation 1 to a *pruning procedure*, which eliminates nodes and weights from the net that contribute to the computation of the result only little or not at all. In such case, a node or weight is selected and eliminated. Then the net is retrained for 100 epochs. If net performance does not drop, the elimination is confirmed; otherwise the deleted node or weight is restored. This procedure is repeated until no further reduction in the size of the net is possible without worsening the performance. [4]

This procedure leads to smaller nets that are easier to analyze and furthermore can reduce the dimensionality of the input data. They have a lower number of weights (i.e. a lower number of free parameters. In the case analyzed here the number of weights was reduced from 649 for the full net to 26 for the pruned net). The weights of a generic example of a pruned network trained on our data are shown in Table 4. There are no weights connecting the input nodes 3, 4, 5, 6, 11, 13, 18, 19, 20, 23 (see Table 4; the meanings of the nodes can be found in Table 3). This means that not all input factors resp. not all

[2] An output value between 0.4 and 0.6 is considered unclassified. However this did not occur in the simulations presented here. Of course the target values are zero and one respectively. Yet, for technical reasons it is preferable to admit a small deviation of the output value from the target values.

[3] Learning parameter is set to 0.2. No momentum. Weights were jogged every epoch by maximally 0.1%. Input patterns are shuffled. The simulations are run on the SNNS network simulator (http://www-ra.informatik.uni-tuebingen.de/SNNS).

[4] More precisely, first we apply the *non-contributing units* algorithm (Dow and Sietsma, 1991), and then pruning of the minimal weight.

their values are relevant for computing the output. Also, all but two hidden nodes have been pruned. So the two remaining suffice to model the interaction between the input factors.

Some nodes have a direct influence on the output node (27), e.g. the node indicating that the rhetorical antecedent was a possessor (node 9). Others influence the outcome only indirectly by interacting with other nodes, e.g. Paragraph Distance (node 24), while yet others influence the output both directly and indirectly. Some nodes enter in multiple ways that seem to cancel each other, e.g. node 14.

Table 4: Weights of a typical pruned net.

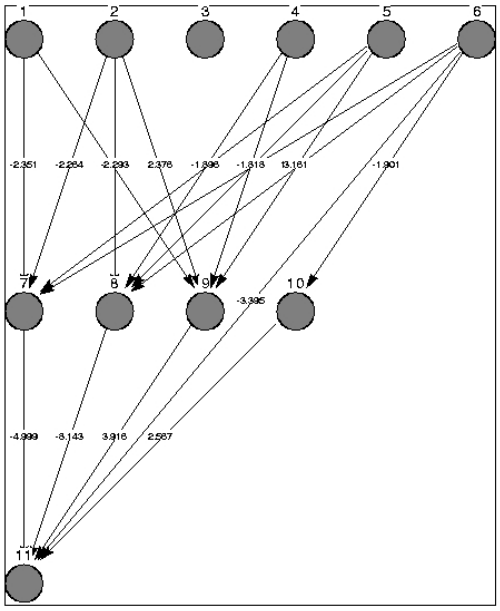
Nodes 1—24 denote the input nodes, 25 and 26 are the two remaining hidden nodes and 27 is the output node. The weights connecting a source and a target node are given in parentheses after the source node.

Target node	Source Nodes							
25	1 (-2.4)	2 (2.1)	8 (-1.7)	12 (1.9)	14 (-1.6)	16 (-2.4)	22 (-4.7)	24 (-4.9)
26	7 (1.7)	10 (-2.0)	12 (-5.0)	14(-1.9)	15 (2.8)	16 (-1.8)	21 (-4.2)	
27	2 (-3.7)	8 (3.9)	9 (2.0)	15 (2.7)	17 (1.8)	22 (-22.0)	25 (10.9)	26 (-10.0)

4.3. Simulation 3 – reduced data set

Pruning again is a partly stochastic procedure, so we repeated the experiment until we got an impression on which factors are almost invariably included. It turned out that Subject and Possessor roles, Protagonisthood, Subjecthood of the antecedent and Type of antecedent are most important, and those nodes related to the rhetorical antecedent are more involved than those for the linear one. As well, the most important distance is Rhetorical Distance.

Thus we use the result of our pruning case study as a hint on how to reduce the dimensionality of the input data. In a third case study we trained a similar net with 12 hidden nodes on a reduced set of only five input factors (corresponding to six input nodes): We included the values Subject and Possessor for the Syntactic Role (nodes 1, 2), Protagonisthood (node 3), whether the rhetorical antecedent was a Subject (node 4), whether it was realized as a pronoun or FNP (node 5), and Rhetorical Distance (node 6). The new net had 12 hidden nodes, corresponding to 103 weights. On this reduced net, we executed the back-propagation learning algorithm for 500 epochs and then pruning (50 epochs retraining for each pruning step) with the same parameters as before. We ended up with a small net (23 parameters), shown in Figure 1, that classified only 8 out of 102 items wrongly. Note that all remaining factors interact strongly, except for protagonisthood (node 3), which has been pruned away.



4.4. Simulation 4 – cheap data set

Reliable automatic annotators for Rhetorical Distance and consequently for all factors related to the rhetorical antecedent, as well as for Protagonisthood, are not available. As these factors require comprehension of the contents of the text, they must be annotated by human experts and are therefore costly. So we decided to replace the rhetorical factors included in Simulation 3 by the corresponding linear ones and Protagonisthood by Animacy. Keeping the six input nodes as before, we added a seventh one to indicate that the linear antecedent was a Possessor and an eighth for Paragraph Distance to help the net to overcome the smaller amount of information that is contained in the linear antecedent factors. Training and pruning proceeded as before.

Figure 1: Net from Simulation 3. The circles denote the nodes, the arrows the weights connecting the nodes, to which the weight strength is added as a real number. Nodes 1—6 are input nodes, 7—10 the nodes in the hidden layer, and node 11 is the output.

One typical resulting network in this case had 32 degrees of freedom. Again Animacy, which had been substituted for Protagonisthood, is disconnected from the rest of the net. On the 102 data items the net produced only six errors (three are due to referential conflict).

5. Comparison to the calculative approach

In Kibrik (1999), the referential choice was modeled by 11 factors using 32 free parameters, that is, the numerical activation of each factor contributed to the activation score. The activation score allowed a prediction of the referential choice in five categories. In our study with neural networks, we modeled only a binary decision (pronoun, FNP) and lifted the requirement of cognitive adequacy. The smallest net in the study, in simulation 3, had only 23 free parameters, 5 input factors, and the best net on the full set of input factors, in Simulations 1 and 2, misclassified only four items, having 26 free parameters.

6. Comparison to Strube and Wolters 2000

Strube and Wolters (2000) use a similar list of factors as Kibrik (1999), except that the costly factors related to the rhetorical antecedent are missing. They analyze a large corpus with several thousand of referring expressions for the categorical decision (FNP, pronoun) using logistic regression. The logistic regression is a form of linear regression adapted for a binary decision.

Thus factor interaction and non-linear relations are not accounted for in their model and they present no cognitive interpretation of their model either.

7. Conclusion and outlook

This is a pilot study testing whether artificial neural networks are suitable to process our data. We trained feed-forward networks on a small set of data. The results show that the nets are able to classify the data almost correctly with resprect to the choice of referential device. A pruning procedure allowed us to single out five factors that still allowed for a relatively good prediction of referential choice. Futhermore, we demonstrated that costly input factors such as Distance to the Rhetorical Antecedent could be replaced by those related to the linear antecedent, which can be more easily collected from a large corpus.

Because of the small amount of data for this pilot study, the result must be taken with due care. But these results encourage us to further develop this approach.

Future work will include a study of a larger data set. This is necessary since neural networks as well as classical stochastics need a large amount of data to produce reliable results that are free of artifacts due to a small corpus. In our corpus, some situations (i.e. an antecedent that is an indirect object) appear only once, so that no generalization can be made. In a larger study the advantages of the neural networks approach can be used fully.

We also aim at reintroducing a cognitive interpretation at a later stage, and want to work with different network types, that not only allow dimensional reduction and data learning, but also an easy way to extract the knowledge of the net in terms of symbolic rules (see e.g. Hammer et al., 2001).

Furthermore, we feel the need not only to model a binary decision (FNP/pronoun), but also to have a more fine-grained analysis. Kibrik (1999) has done the first step in this direction, allowing for five different categories that not only state that a pronoun or FNP is expected, but also to what degree a FNP in a particular situation can be replaced by a pronoun and vice versa.

We suggest a statistical interpretation of referential choice in the following sense: if a human expert judges that a particular FNP could be replaced by a pronoun, s/he must have experienced that in a very similar situation where the writer indeed had decided to realize the other alternative. The expert will be more certain that substitution is suitable if s/he has often experienced the alternative situation. Thus we think it is promising to replace Kibrik's five categories by a continuous result variable that ranges from zero to one and is interpreted as the probability that referential choice realizes a pronoun in the actual situation: one means a pronoun with certainty, zero means a FNP with certainty, and 0.7 means that in 70% a pronoun is realized and a FNP in the remaining 30%.

Acknowledgements

Andrej Kibrik's research has been supported by grants 03-06-80241 and 02-06-80037 of the Russian Foundation for Basic Research. I also express my gratitude to the Alexander von Humboldt Foundation and Max-Planck-Institute for Evolutionary Anthropology that made my research in Germany (2000-2001) possible.

References

1. Chafe, W., 1994. Discourse, Consciousness, and Time. The Flow and Displacement of Conscious Experience in Speaking and Writing. Chicago: University of Chicago Press.
2. Dale, Robert. 1992. Generating referring expressions. Cambridge, Mass.: MIT Press
3. Dow, R., Sietsma, J., 1992. Creating Artificial Neural Networks that Generalize. Neural Networks 4(1): 67–79.
4. Fine, T.L., 1999. Feedforward Neural Network Methodology. New York: Springer.
5. Fox, B., 1987. Discourse Structure and Anaphora. Cambridge: Cambridge University Press.
6. Hammer, B., A. Rechtien, M. Strickert, T. Villmann. Vector Quantization with Rule Extraction for Mixed Domain Data. Submitted to ILP 2002.
7. Kibrik, A.A., 1996. Anaphora in Russian narrative prose: A cognitive account. In B. Fox (ed.), Studies in Anaphora. Amsterdam and Philadelphia: John Benjamins.
8. Kibrik, A.A., 1999. Reference and working memory: Cognitive inferences from discourse observation. In K. van Hoek, A.A. Kibrik, and L. Noordman (eds.), Discourse Studies in Cognitive Linguistics. Amsterdam and Philadelphia: John Benjamins.
9. Kibrik, A.A. 2000. A cognitive calculative approach towards discourse anaphora. In P. Baker, A. Hardie, T. McEnery and A. Siewierska (eds.), Proceedings of the Discourse anaphora and anaphor resolution conference (DAARC2000). Lancaster: University Centre for Computer Corpus Research on Language.

10. Strube M., M. Wolters, 2000. A Probabilistic Genre-Independent Model of Pronominalization. NAACL '00, 18–25.
11. Tomlin, R. and M. Pu, 1991. The management of reference in Mandarin discourse. Cognitive Linguistics 2: 65–93.