NON-CATEGORICAL REFERENTIAL CHOICE

Mariya V. Khudyakova
National Research University Higher School of Economics
mariya.kh@gmail.com

Andrej A. Kibrik
Institute of Linguistics RAN & Lomonosov Moscow State University
aakibrik@gmail.com

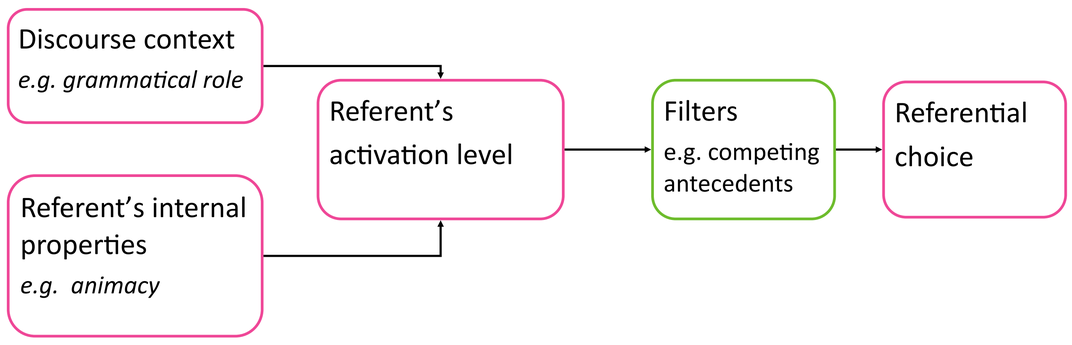
Grigory B. Dobrov
Trafika
wslcdg@gmail.com

REFERENTIAL CHOICE

Choice between forms of referential expressions at a given moment of discourse:

John / The man / This man / He came

Multi-factorial model



NON-CATEGORICAL REFERENTIAL CHOICE

$$\text{Accuracy of modeling} = \frac{\text{Amount of properly predicted referential forms}}{\text{Total amount of referential expressions}}$$

Properly predicted forms = ?

In natural discourse there is usually a subset of instances in which two or more referential expressions are equally appropriate.

REFRHET3 CORPUS

Based on the RST Discourse Treebank (Carlson, Marcu & Okurowski, 2003)
(contains annotation for rhetorical structure)

Texts from Wall Street Journal

MoRA (Moscow Reference Annotation) Scheme: annotation for

- Coreference
- Potential activation factors

64 texts

1852 anaphor-antecedent pairs

MODELING

Two-way choice: pronouns vs. full NPs

Three-way choice: pronouns vs. descriptions vs. proper names

WEKA package (Frank et al. 2010)

Method	Accuracy for two-way task	Accuracy for three-way task
Logistic regression	87.2%	71.3%
Decision trees C4.5	93.7%	74.0%
Decision Trees C4.5 + boosting	89.4%	76.1%
Decision Trees C4.5 + bagging	89.5%	74.0%

EXPERIMENT

Material: 27 texts from the corpus, 31 "problem points" (choice of Decision trees algorithm differs from the original text)

Participants: 47 students, aged 18-21, speaking English on Expert level

Task: Please choose all appropriate options (it is possible to choose more than one option).

Task example:

First Tennessee National Corp. said it would take a $ 4 million charge in the fourth quarter, as a result of plans to expand its systems operation. The banking company said it reached an agreement in principle with International Business Machines Corp. on a systems operations contract calling for IBM to operate (First Tennessee's / the company's / its) computer and telecommunications functions. Further, under (the agreement / it), First Tennesse would continue to develop the software that creates customer products and sevices. Because personal computers will soon be on the desks of all of our tellers, and customer service and loan representatives, information will be instantly available to help customers with product decisions and provide (the customers / them) with information about their accounts, according John Kelley, executive vice president and corporate services group manager at First Tennessee. <…>

RESULTS

Problem point type (original text—algorithm choice)	Human choice:	
	As in the original text	As predicted by the algorithm
Description—pronoun	67%	33%
Proper name—pronoun	61%	39%
Pronoun—description	53%	47%
Pronoun—proper name	55%	45%

REFERENCES

Carlson L., Marcu D., Okurowski M. E. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. Springer Netherlands, 2003.
Frank E., Hall M., Holmes G., Kirkby R., Pfahringer B., Witten I. H., Trigg L. 2010. Weka-a machine learning workbench for data mining. In: Data Mining and Knowledge Discovery Handbook. Springer US, 1269-1277.