INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 12(2):217-226, April 2017.

Detecting Bridge Anaphora

D. Gîfu, L.I. Cioca

Daniela Gîfu*
1. Romanian Academy - Iaşi branch
Codrescu, 2 Iaşi, 700481, Romania
2. "Alexandru Ioan Cuza" University of Iaşi
General Berthelot St., 16, Iaşi, 700483, România
"Alexandru Ioan Cuza" University of Iaşi, România
*Corresponding author: daniela.gifu@info.uaic.ro

Lucian-Ionel Cioca
"Lucian Blaga" University of Sibiu
10, Victoriei Bd., Sibiu, 550024, România
lucian.cioca@ulbsibiu.ro

Abstract: The paper presents one of most important issues in natural language processing (NLP), namely the automated recognition of semantic relations (in this case, bridge anaphora). In this sense, we propose to recognize automatically, as accurately as possible, this type of relations in a literary corpus (the novel *Quo Vadis*), knowing that the diversity and complexity of relations between entities is impressive. Furthermore, we defined and classified the bridge anaphora type relations based on annotation conventions. In order to achieve the main goal, we developed a computational instrument, BAT (*Bridge Anaphora Tool*), currently still in a test (and implicitly an improvable) version. This study is intended to help especially specialists and researchers in the field of natural language processing, linguists, but not only.
Keywords: bridge anaphora, annotated novel, Bridge Anaphora Tool, testing corpus, corpus-driven, statistics.

1 Introduction

The novelty of this study consists in the development of a web application for the automated identification of bridge anaphora type relations in a corpus from the literary area. In this case, the target is the Romanian version of the novel *Quo Vadis*, authored by the Nobel laureate Henryk Sienkiewicz [24].

Initially, a similar study carried out by the same team consisted in the supervised extraction of Bridge Anaphora type relations, using WEKA statistics [8]. Moreover, there was defined a set of annotation conventions for 11 bridge relations as a result of manual annotations made by a team of trained students in Computational Linguistics.

The hypothesis of this paper is that the triggers have a fundamental role in the automated recognition of semantic relations generally and particularly of bridge anaphora relations.

The paper is structured in 5 sections. After a brief introduction about the importance of this study, section 2 mentions some important works focused on bridging anaphora. Section 3 describes bridge anaphora relations in the context of semantic relations and section 4 describes a new tool functionality, called BAT (*Bridge Anpahora Tool*). The last section highlights conclusions and mentions the future intentions, one of the main projects of Romanian researchers in NLP.

2 State of the art

In our context, the semantic relations play a fundamental role in the information extraction process [1], regardless of the nature of the corpus [2, 9, 10]. Up to now, researchers in the NLP area have allocated a lot of time to identifying the best annotation conventions of semantic relations for various literary types [3, 4, 12] based on which the process of automated recognition of semantic relations was not only simplified, but the accuracy of results increased as well, for example in their unsupervised extraction [7].

One of the best known studies on the "bridging" concept originates with H.C. Herbert [11]. He starts from several scenarios in which an inference step is needed in order to understand the sense intended by the speaker and he states that the text itself does not offer the solution for solving the inference relation; the reader (or the computational instrument/machine) must use his/its knowledge on the anaphora and the antecedent in order to make a correct text interpretation. In the automated recognition of semantic relations, a special attention is granted to the anaphora resolution [23, 25], using statistic models [19, 22], something that we too exploited.

NLP uses for recognizing entities and identifying relations in the text (bridging) systems based on manually created rules (see Hobb's algorithm) [15], but also systems using statistical models that are in turn based on automated learning techniques in order to lessen the workload, models such as Conditional Random Fields (CRF) [26].

3 Bridge anaphora in semantic relations context

In order to better understand a content, we need thinking instruments, necessary for discovering new ideas or for clarifying the existing ones, illustrating the link between them. The semantic relations [16] describe these interactions, that are indispensable for interpreting texts. The properties of semantic relations were described in [17], this marking the relations between two entities (called poles) as open class. The application describes 10 types of bridge anaphora[1].

3.1. A short introduction about semantic relations

The semantic relations are represented as being distributions over several paragraphs [18]. In processing the natural language, the semantic relations play a fundamental role in the field of Information Extraction (IE), that targets the automated extraction of structured information referring to entities such as person names, localities etc. from semi-structured or unstructured texts.

The ability to identify and understand these relations in a text can b useful in very many directions, such as: Machine Translation - MT; Computer Assisted Assessment - CAA; Clustering and so on.

In order to create an instrument that can carry out, for example the automated translation, the interpretation of anaphora is also very important, especially in cases in which the translation is from a language in which the pronouns have different forms for each gender, into a language in which the pronoun has the same form regardless of gender [15, 22].

[1] **Class-of** - relation between PERSON-CLASS & PERSON; **Has-as-member** - relation between PERSON-GROUP & PERSON; **Has-as-part** - relation between PERSON & PERSON-PART; **Has-as-subgroup** - relation between PERSON-GROUP & PERSON-GROUP; **Has-name** - relation between PERSON & PERSON-NAME; **ISA** - relation between PERSON & PERSON-CLASS; **Member-of** - relation between PERSON & PERSON-GROUP; **Name-of** - relation between PERSON-NAME & PERSON; **Part-of** - relation between PERSON-PART & PERSON; **Subgroup-of** - relation between PERSON-GROUP & PERSON-GROUP.

3.2. About bridge anaphora

Bridge anaphora [8] are referential semantic relations (beneath the co-referential or anaphoral ones) [5, 6] that includ linguistic expressions that give meaning to the analysed text (here, the narrative "thread" of the novel). Our documentation shows that the analysis of semantic relations is focused on structured corpuses such as: online newspapers, blogs, Wikipedia texts etc. [1].

A bridge anafora or "bridging" is a semantic relation that represents a link between the anaphora and the antecedent [11, 12]. These two elements will be mentioned in the following also as poles of a bridge-type semantic relation. In the next section we present the 10 types of bridge anaphora relations based on which the BAT was developed.

An example of bridge-type semantic relation:

Andrei este numit în diferite cercuri micuţul, din cauza înălţimii.

—>(En.) Andrei is called in different circles the little guy, because his height.
where:

- *Andrei* is an antecedent;
- *micuţul* —>(En.) *the little guy* is an anafor.

3.3. Bridge anaphora vs. anaphora

A bridge anaphora type relation differs from an anaphorical relation firstly by the fact that it can be identified in the text using a *trigger*. This trigger can be a word or a group of words that has the property of indicating the presence in the text of the bridge anaphora relation, helping to identify it.

In the following, we will exemplify the anaphorical relation and the bridge anaphora type relation in oder to clarify the difference between the two relations, both being referential type relations:

- Anaphorical relation (coreferential)

1:[Marcus] era foarte supărat pentru toate cele întâmplate în ultima perioadă, însă 2:[el] nu avea de gând să renunţe. —>(En.) 1:[Marcus] was very upset about what happened lately, but 2:[he] was not going to give up.
=>[2] anaphorical relation [1];

- Bridge anaphora type relation (below, the type **class-of**[2])

Cândva, 1:[Petronius] fusese guvernator în 2:[Bitinia]... —>(En.) Sometime 1:[Petronius] was governor in 2:[Bithynia]...
=>[1] bridge anaphora type relation [2], while governor in is the trigger for this relation.

This is a segmentation annotation in XML standoff format:

```
<W LEMMA="cândva" MSD="Rg" POS="ADVERB" id="1" offset="0">Cândva</W>
<W LEMMA="," MSD="COMMA" id="2" offset="6">,</W>
<CLAUSE CONTINUE="27" ID="CLAUSE31">
<ENTITY ID="E000900036" TYPE="PERSON">
```

[2]**Class-of** - is a bridge anaphora type relation linking a PERSON-CLASS type concept to a PERSON type instance.

```
<REFERENTIAL FROM="E000900036" ID="REF000900582" TO="E000700030" TYPE="coref">
<W Case="oblique" Definiteness="no" EXTRA="NotInDict" Gender="feminine"
LEMMA="Petronius" MSD="Npfpon" Number="plural" POS="NOUN" Type="proper" id="3"
offset="8">Petronius</W>
</REFERENTIAL>
</ENTITY>
</CLAUSE>
<W EXTRA="intranzitiv" LEMMA="fi" MSD="Vmil3s" Mood="indicative"
Number="singular" POS="VERB" Person="third" Tense="long" Type="predicative"
id="4" offset="18">fusese</W>
<ENTITY ID="E000900037" TYPE="PERSON">
<W Case="direct" Definiteness="no" Gender="masculine" LEMMA="guvernator"
MSD="Ncmsrn" Number="singular" POS="NOUN" Type="common" id="5"
offset="25">guvernator</W>
</ENTITY>
<W LEMMA="în" MSD="Sp" POS="ADPOSITION" id="6" offset="36">în</W>
<CLAUSE CONTINUE="31" ID="CLAUSE32">
<ENTITY ID="E000900038" TYPE="LOCATION">
<REFERENTIAL FROM="E000900038" ID="REF000900584" TO="E000900036"
TYPE="class-of">
<REFERENTIAL FROM="E000900038" ID="REF000900584" TO="E000800035" TYPE="coref">
<W EXTRA="NotInDict" LEMMA="Bitinia" MSD="Np" POS="NOUN" Type="proper" id="7"
offset="39">Bitinia</W>
</REFERENTIAL>
</REFERENTIAL>
</ENTITY>
</CLAUSE>
```

The anaphorical relations are a widely debated subject [12, 13, 14], proven by numerous specialty papers that present computational instruments for the automated identification of these relations, especially for the pronominal anaphora [15, 22]. This type of relation is much easier to identify in the text, as opposed to a bridge anaphora type semantic relation, because both poles of the the relation, the anaphora and the antecedent refer to the same entity [20]. In order to be able to automatically identify bridge type anaphorical relations, there is necessary a more complex mechanism, that would carry aut in a first phase a preprocessing of the text for its de-ambiguization that consists in segmentation, tokenization, lemmatization, part-of-speech tagging, name entity recognition, and anaphora resolution.

4 BAT - description

Bridge Anaphora Tool is a computational instrument implemented in Java language, on the framework Java Server Faces and uses a series of libraries[3]. BAT is created for the automated recognition of bridge type semantic relations, more precisely of the 10 types of referential relations for which annotation conventions have been determined.

The output XML file was used in the process of training and testing. We chose the novel Quo Vadis [24], given that it is a corpus translated into more than 40 de languages, having an impressive number of entities and semantic relations. Using the instrument PALinkA [21]

[3]see http://primefaces.org/

the entities and semantic relations were annotated manually. The annotator was already used successfully for annotating the novel Quo Vadis, a work presented in [3].

This web application (fig. 1) executes in a first phase the training process after which the automated recognition can be initiated.

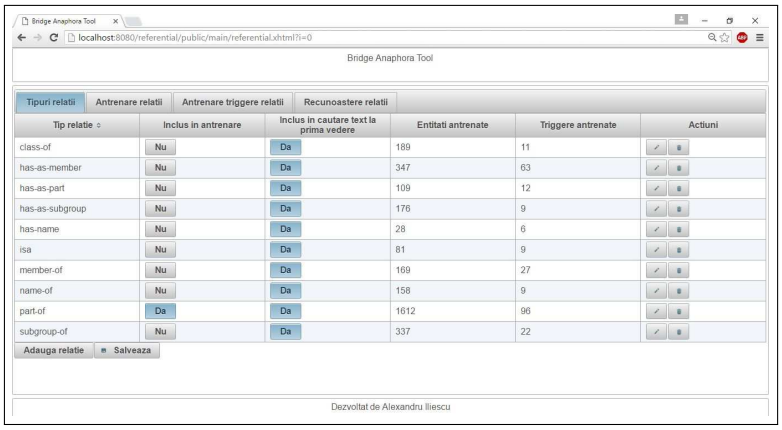

Figure 1: The interface of the computational tool

In the following, we describe briefly the work methodology. For the trening process, following steps were conceived:

- The option "YES" is selected for the relations that will be included in the training;

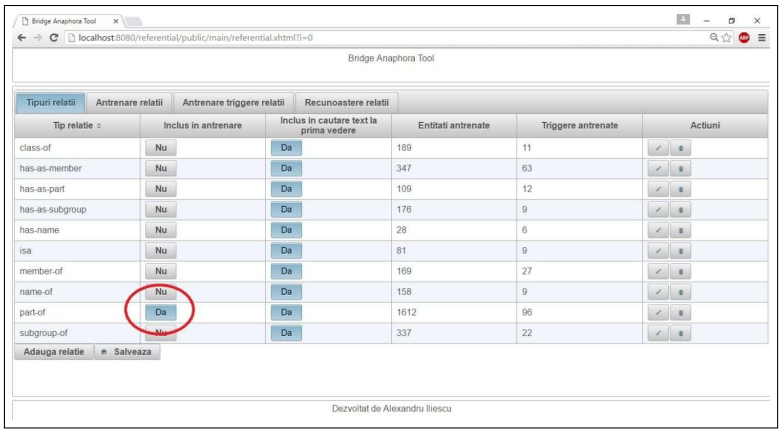

Figure 2: BAT - selecting relations for training

- The XML file is loaded from the application, using the button "Train relations", the XML is selected (the manually annotated corpus) after which the button "Process file" is pressed.

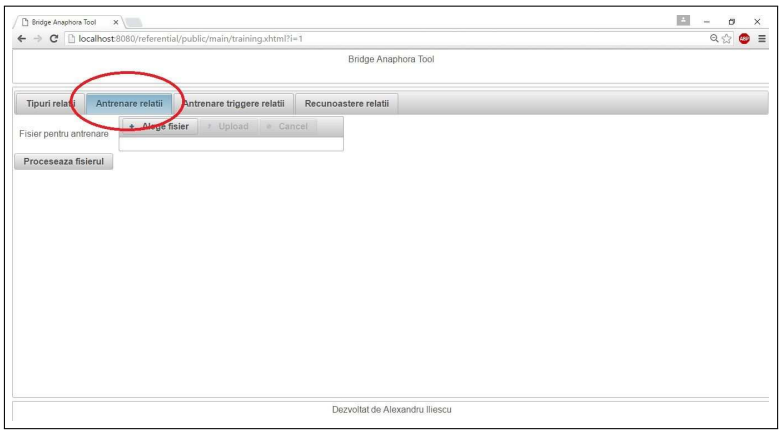

Figure 3: BAT - menu for introducing the corpus and for trening relations

When the BAT identifies the tag <REFERENTIAL>, it carries out four steps:
- it saves in the MySQL database the type of relation (it can be one of the 10 "class-of", "has-as-member", etc.) in the table "referential_type";

Figure 4: BAT - the table "referential_type" generated after the training

- it saves the type of entity from the identified relation (in the example above we have TYPE="PERSON") in the table "entity_type";

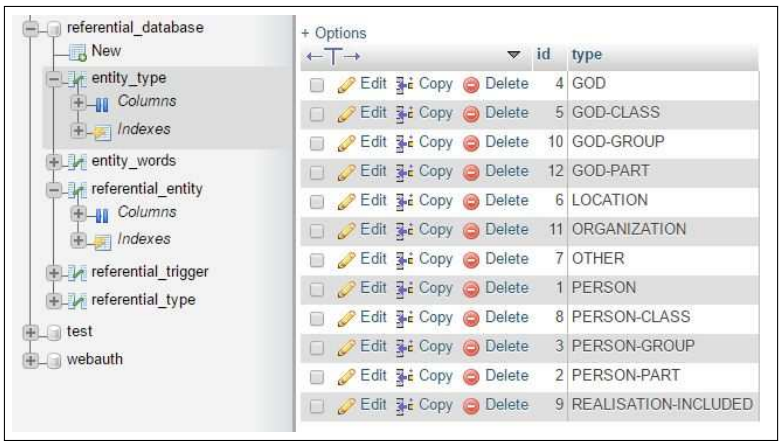

Figure 5: BAT - the table "entity_type" generated after the training

- it saves the word/words from the tag <ENTITY>specific to the relation in the table "entity_words"; if there are two or more words, they are concatenated with the symbol "|";

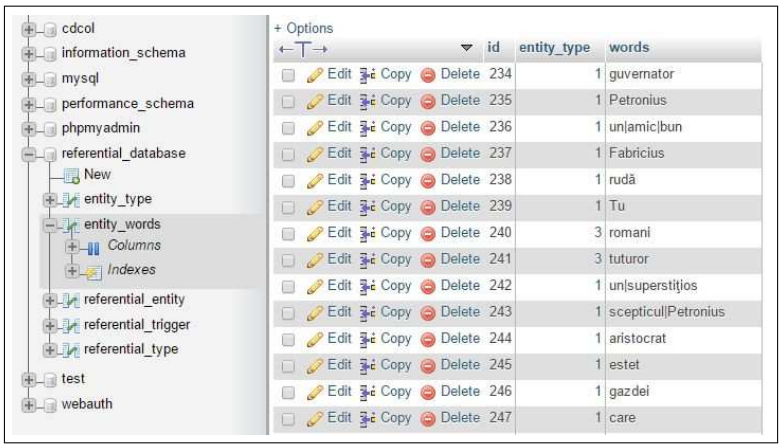

Figure 6: BAT - the table "entity_words" generated after training

- it saves the actual structure of a bridge type semantic relation, i.e. the "TYPE" and the words forming it, in XML they being identifiable with the elements "ID", "FROM" and "TO".

Figure 7: BAT - the table "referential_entity" generated after training

The processing of the XML file in the training phase of the BAT for one or more bridge type semantic relations can take from one minute to several hours, function of the number of relations existing in the annotated corpus. For the "part-of" relation, the training took 2.67 hours, being the most often encountered in the XML file, with a number of 1612 relations.

5 Statistics and interpretations

Bridge Anaphora Tool used for training 66% of the corpus of Quo Vadis.

Mitkov (1998) suggested, for measuring the performance of a computational instrument aiming at identifying anaphorical relations in the text, an equation that defines its success rate.

The definition of the success rate is as follows:

BAT success rate = 534 correctly identified relations / 1035 total existing relations = 61.5%.

So at this moment, the BAT recognized correctly over 61% of the bridge anaphora type semantic relations that should have been identifying in the text, thus fulfilling the set goal.

Table 1: The results of recognizing the bridge type semantic relations with BAT

Bridge Anaphora types	Bridge Anaphora number identified with BAT in driven corpus	Bridge Anaphora number identified automatically in testing corpus	Bridge Anaphora number identified manually in testing corpus
class-of	189	58	28
has-as-member	347	115	82
has-as-part	109	31	12
has-as-subgroup	176	55	22
has-name	28	9	7
isa	81	25	14
member-of	169	51	38
name-of	158	53	29
part-of	1612	530	249
subgroup-of	337	108	53
Total	**3206**	**1035**	**534**

We think that the variations of the values in the column "number of relations identified automatically by BAT in the testing corpus" are due also to the fact that the instrument searches "mechanically" in the preprocessed text rigid definitions of the relations.

For example: entity of the type PERSON-NAME + PERSON =>relation "name-of".

Moreover, there exist two relations that have the same definition, namely the relations: *has-as-subgroup* and *subgroup-of* being given by the entities of type PERSON-GROUP+PERSON-GROUP, the only difference between them being made by the triggers, during testing.

Conclusions and future work

This paper presents a methodology for the automated recognizing of 10 bridge anaphora (or bridging) type semantic relations, each having several particularities. The achieved results are promising, offering a base for future researches. We suggest using in parallel of machine learning models (Naïve Bayes and Support Vector Machines).

The BAT instrument, developed for the automated recognition of Bridge Anaphora relations, will be improved through the addition of several triggers to the existing list, or in the situation in which there would be available even more data for training.

BAT is far from being a perfect instrument, but it can be improved since it showed to be efficient at least for an experimental purpose for various applications in the NLP area.

Acknowledgments

We thank Alexandru Iliescu for developing the BAT instrument. We are also grateful to all colleagues from NLP-Group@UAIC-FII who developed the tools for natural language processing used in this research.

Bibliography

[1] Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., Etzioni, O. (2007); Open information extraction from the Web, *IJCAI07 Proceedings of the 20th international joint conference on Artifical intelligence*, 2670-2676.

[2] Buraga, S.C., Cioca, M., Cioca, A. (2007); Grid-based decision support system used in disaster management, *Studies in Informatics and Control*, 16(3):283-296.

[3] Cristea, D., Gîfu, D., Diac, P., Maraunduc, C., Bibiri, A., Scutelnicu, A., Colhon, M. (2014); *Quo Vadis: A Corpus of Entities and Relations*, Springer International Publishing Switzerland, 2014.

[4] Cristea, D., Dima, G. E., Postolache, O. D., Mitkov, R. (2002); Handling complex anaphora resolution cases, *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium, Lisbon, 2002*, 1-6.

[5] Branco, A., McEnery, T., Mitkov, R. (2005); *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, John Benjamins, 2005.

[6] Cojocaru, L. (2000); The study of the anaphoric relationship on the Romanian corpora, Inference in Computational Semantics, at the Inference in Computational Semantics, *ICoS-2, Schloss Dagstuhl, Germany, July 29-30.*

[7] Conrath, J, Afantenos, S., Asher, N., Muller, P. (2014); Unsupervised extraction of semantic relations using discourse cues, *International Conference on Computational Linguistics - COLING (Dublin, Ireland)*, 2184-2194.

[8] Gîfu, D., Iliescu, A. (2014); Analysis of Bridge Anaphora across novel, *Procedia- Social and Behavioral Sciences*, 180: 1474-1480.

[9] Gîfu, D.; Cioca, M. (2013); Online civic identity. Extraction of features, *Procedia - Social and Behavioral Sciences*, 76:366-371.

[10] Gîfu, D.; Cioca, M. (2014) Detecting Emotions in Comments on Forums, *International Journal of Computers Communications & Control*, 9(6):694-702.

[11] Herbert, H. C. (1975); Bridging. *Proceedings of the 1975 Workshop on. Theoretical Issues in Natural Language. Processing, TINLAP '75*, 169-174.

[12] Hendrickx, I., Clercq, O., Hoste, V. (2011); Analysis and reference resolution of bridge anaphora across different text genres, *Anaphora Processing and Applications - 8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2011, Faro, Portugal*, LNAI 7099:1-11.

[13] Krahmer, E., Piwek, P. (1997); Varieties of anaphora, *Proceedings of the 11th Amsterdam Colloquium, University of Amsterdam*, 5-20.

[14] Korzen, I., Buch-Kromann, M. (2006); Anaphoric relations ĂŽn the Copenhagen Dependency Treebanks, *Proceedings of COLING-ACL 06*, 3:83-98.

[15] Lappin, S., Leass, H. J. (1994); An Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics*, 20(4):535-561.

[16] Liang, T., Wu, D. (2010); Automatic Pronominal Anaphora Resolution ĂŽn English Texts, *Computational Linguistics and Chinese Language Processing*, 9(1):21-40.

[17] Murphy, M. L. (2003); *Semantic relations and the lexicon Antonymy, synonymy, and other paradigms*, Cambridge University Press, Cambridge, UK, 2003.

[18] Năstase, V., Nakov, P., Seaghdha, D. O., Szpakowicz, S. (2013); *Semantic relations between nominals*, California: Morgan & Claypool Publishers, 2013.

[19] Niyu, G., Hale, J., Charniak, E. (1998); A Statistical Approach to Anaphora Resolution, *Proceedings of the Sixth Workshop on Very Large Corpora, Montreal, Canada*, 161-170.

[20] Nedoluzhko, A., Mirovsky, J., Ocelak, R., Pergler, J. (2009); Extended Coreferential Relations and Bridging Anaphora in the Prague Dependency Treebank, *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India*, 1-16.

[21] Orăşanu, C. (2003); PALinkA: A highly customisable tool for discourse annotation, *Proceedings of the 4 th SIGdial Workshop on Discourse and Dialogue, ACL'03*, 1-5.

[22] Rello, L., Ilisei, I. (2009); A comparative study of spanish zero pronoun distribution, *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages (ISMTCL)*, 1-5.

[23] Wasow, T. (1967); *Anaphoric relations in English*, Ph.D. dissertation, Massachusetts Institute of Technology, 1967.

[24] Sienkiewicz, H. (1991); *Quo Vadis*, translated in Romanian by Luca, R. and Linţă, E., Ed. Tezi, Bucharest.

[25] Singh, S., Lakhmani, P., Mathur, P., Morwal, S. (2014); Analysis of Anaphora Resolution System for English Language, *International Journal on Information Theory*, 3(2):5157, DOI : 10.5121/ijit.2014.3205.

[26] Žitnik, S, Šubelj, L, Bajec, M (2014); SkipCor: Skip-Mention Coreference Resolution Using Linear-Chain Conditional Random Fields, *PLoS ONE*, 9(6): e100101. doi:10.1371/journal.pone.0100101.