Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2017”

Moscow, May 31—June 3, 2017

BRIDGING ANAPHORA RESOLUTION FOR THE RUSSIAN LANGUAGE

Roitberg A. M. (cvi@yandex.ru)[1,2],
Khachko D. V. (mordol@lpm.org.ru)[1]

[1]IMPB RAS- Branch of KIAM RAS, Puschino, Russia;
[2]School of Linguistics HSE RSU, Moscow, Russia

Presented in this report are the initial findings of automatic bridging anaphora recognition and resolution for the Russian language. For a resolution of F-measure = 0.65 we use a manually-annotated bridging corpus and machine-learning techniques to develop a classifier to predict bridging anaphors, bridging anchors, and bridging pairs. In addition to this, we discuss the features used for the classifier and discuss the importance of each feature. Experimental results show that our classifier works well, however, potential improvements can be made, these improvements will be explored.

Key words: bridging anaphora, bridging anaphora recognition, bridging anaphora resolution

РАЗРЕШЕНИЕ БРИДЖИНГ АНАФОРЫ НА МАТЕРИАЛЕ РУССКОГО ЯЗЫКА

Ройтберг А. М. (cvi@yandex.ru)[1,2],
Хачко Д. В. (mordol@lpm.org.ru)[1]

[1]ИМПБ РАН—филиал ИПМ им. М. В. Келдыша РАН),
Пущино, Россия; [2]Школа лингвистики НИУ ВШЭ

В статье представлены первые результаты автоматического распознавания бриджинг-анафоры для русского языка. Для распознавания анафоры F-мера равна 0,65. Распознавание бриджинг-анафоры проводилось с применением методов машинного обучения. Для обучения классификатора мы использовали корпус коротких новостных текстов на русском языке с ручной разметкой бриджинг-анафоры. В данной работе обсуждаются наборы признаков, которые использовались для обучения, а также значимость каждого из признаков.

Ключевые слова: бриджинг-анафора, распознавание бриджинг-анафоры, разрешение бриджинг-анафоры

1. Introduction

Bridging anaphora, cover a broad class of semantic and discourse relations and play an important role in text cohesion. Therefore, bridging anaphora recognition and resolution is of vital importance for a variety of different NLP tasks.

Thus far, automatic bridging anaphora resolution has been in its early stages of development. Previous studies on bridging resolution include: [Poesio et al. 2004, Fan et al 2005] which uses a semantic approach to bridging resolution, they restrict bridging relations to several semantic relations, like part-of, set-subset etc. [How et al. 2013; Sasano 2009] considers using probabilistic methods, finally, in the last paper, bridging anaphora is resolved just as zero anaphora is. We too use a probabilistic approach for bridging anaphora recognition and resolution. Currently, there are no previous studies on bridging anaphora for the Russian language; our goal is to bridge this gap.

Following [Clark 1975], we define bridging as all anaphoric relations between non-coreferential entities. Clarks definition is very general; all the works on bridging anaphora narrow this definition in some way. For example; [Lüngen 2008; Poesio 2004] restrict bridging-anaphora semantically, [Hou et al. 2013] restricts all bridging expressions to noun phrases etc.

We apply a completely new syntactic type of restriction to bridging cases (See 2.1). Our non-semantic restriction approach is attributed to the fact that we don't apply WordNet or a similar resource for bridging resolution, as distinguished from [Poesio et al. 2004] for example. Unfortunately, there is no such resource as English WordNet for Russian.

The paper is set out as follows: In Section 2.1 we describe our syntactic-oriented approach—Genitive Bridging; Section 2.2 details the training and testing corpus used throughout the paper—RuGenBridge, Section 3 explores our machine-learning experiments, first describing in general terms the machine learning procedure (3.1), then going on to list the learning features for bridging elements (3.2.1), bridging anchors (3.2.2) and bridging pairs (3.2.3). In Section 4 we present the results and discuss the advantage of our methods and explore future directions of research.

2. Data collection and Preparation

2.1. Genitive Bridging

We applied a specific syntactic-oriented approach to bridging anaphora called 'genitive bridging'. We capture the bridging relation in cases where the anchor NP and the bridging-anaphor NP are: 1) anaphorically linked and 2) the heads of anaphorically linked NPs can form a grammatical genitive construction. The bridging anaphor is the head of genitive construction and the anchor a genitive dependence.

(1) *Tam stoyal <u>gruzovik</u> s naklejkami na **kabine**.*
'*There was a huge <u>truck</u> with stickers on the **cab***'

Example 1: *kabina—gruzovik* [*cab—truck*] is a case of genitive bridging: 1) the entities are anaphorically linked; the cab of the truck was mentioned in the previous sentence; 2) at the same time *kabina gruzovika [cab truck.Gen] 'cabin of the truck'* is a grammatical genitive construction in Russian. So, in such a case, we annotate with genitive bridging: *cab* → *truck*.

We don't restrict bridging relations to some semantic relations. However our observations show that most of the genitive bridging pairs are as follows: 1) part-of relations, as in the example above; 2) political positions—geographic name: *president—USA*; position—organization: principal—school; 3) something located somewhere: *schools—Moscow*; 4) object—its possessor: *flat—landlord*; 5) expressions with names of measures: *oil—barrel*); 6) collocations, mostly deverbative nouns: *rates—increase, robbery—bank*. For further details see [Roitberg, Nedoluzhko 2016]

Note that; among the features used for bridging recognition in [Poesio 2004-B] there is one that can be described in the following way: two expressions are more likely linked to a bridging relation if they frequently appear in syntactic construction *X of Y.* To evaluate it, one must investigate several potential google queries of the form "the NBD of the NPA", where NBD is a head noun of bridging description and NPA is a head noun of a potential bridging antecedent. X of Y is a standard translation for Russian genitive construction X + Y.Gen.

2.2. RuGenBridge Corpus

To train and test our bridging-recognition system, we annotated the Russian corpus, highlighting the genitive bridging—RuGenBridge. It consisted of short news texts (up to 10 sentences) from internet news sources. Currently, we have annotated 339 texts or 61,076 tokens, and have tagged 609 genitive bridging pairs.

Segments of speech and syntactical links were annotated automatically by FreeLing1 and MaltParser2 [Nivre et al 2006] respectively. Bridging relations were annotated manually using BRAT3 tool.

The first part of the corpus (190 texts) was manually annotated by two annotators, with the agreed F-measure = 0.7. The remainder of the text was annotated by 1 annotator.

We annotated genetive bridging relations and coreferential chains for bridging anaphors and anchors. See Example 2.

(2) *Posle vozvrasheniya iz Irana on rassakazal o poezdke v etu stranu. «V Irane ochen' gostepriimnyj **narod**»*
*'After his returning from Iran, he told about the journey to this country. «In Iran there are very welcoming **people**»'*

In Example 2 we annotated the bridging link *narod* → *Iran* ('*people* → *Iran*') and the coreferential chain *Iran—strana—Iran* '*Iran—country—Iran*'. We postulate bridging relations between the bridging anaphor and the whole anchor's coreferential chain, as in the Prague Dependency Bank [Poláková 2013]. We consider two annotations as equal if their bridging anaphors are identical and their anchors belong to the same coreferential chains.

[1] http://nlp.lsi.upc.edu/freeling/

[2] www.maltparser.org

[3] http://brat.nlplab.org

As well as the manually-annotated corpus, we also used a 5 million automatically-part-of-speech-tagged news corpus to train the Word2Vec4 model. Later we use Word2Vec outputs to calculate semantic similarity measures between the nouns of the texts and bridging anaphors or anchors that have been manually annotated.

3. Machine Learning Techniques for Bridging Resolution

For machine learning experiments, we use Python with libraries Pandas5 and Scikit-learn tool6. To reveal cases of bridging anaphora, we use a Random Forest Classifier algorithm because it produces the highest quality results, however, we also conducted some experiments with Logistic Regression and Decision Tree algorithms (see Section 4).

3.1. Procedure

Firstly, we use the whole corpus to calculate TF-IDF for all words in the corpus. Once this is done, we divide our corpus into two unequal parts and use two-step machine learning procedures to train our classifier (See 3.1.1, 3.1.2). We use the larger part (80,000 tokens) called Part 1 bellow for Step 1 and the smaller (14,000 tokens)—Part 2 bellow for Step 2. Step 1 involves training Classifier 1 to recognize potential bridging anaphor/anchors; Step 2 involves training Classifier 2 to recognize bridging pairs. For both steps we apply cross-validation techniques with k-fold = 4. The average was then calculated with the AUC measure also being calculated after each run.

3.1.1. First Part: Step 1

We take all of the bridging anaphors/anchors from Part 1 of the corpus and choose the 10 most semantically similar nouns for each bridging anaphor/anchor (according to Word2Vec).

Once this is done we train the classifier to predict bridging anaphors/anchors. We use an analogous procedure for bridging anaphors and anchors. Let us consider anchors for example: we take all manually annotated anchors as positive examples and add to this “positive” set a group of random nouns as negative examples. The negative set is seven times larger; the best proportion between positive and negative examples was experimentally derived. This data is used for Classifier 1. This classifier was then used to predict bridging anaphors/anchors in the second step.

3.1.2. Second Part: Step 2

As previously mentioned, we use the first step classifier to automatically annotate bridging anaphors and anchors in Part 2. For this task, we optimize Classifier 1 to a very high precision ($P = 0.98 - 1.00$), with such settings it identifies almost all bridging elements/anaphors, moreover, it identifies 10 times more wrong nouns. We then take all of these bridging elements and anchors, match them to Golden standard, and mark

[4] https://code.google.com/archive/p/word2vec/

[5] http://pandas.pydata.org

[6] http://scikit-learn.org/stable/

the real bridging pairs as positive examples and wrong bridging pairs as negative examples. Finally, we use this data to train Classifier 2 to predict bridging pairs.

3.2. Features

We use 2 different feature sets for the bridging anaphors/anchors classifier (Classifier 1) and bridging-pairs classifier (Classifier 2).

3.2.1. Step 1. Feachors for Anchors Prediction

We used eight features to train a classifier for anchor recognition and prediction. These features include:

1. *Semantic similarity to anchor anaphor*—as previously mentioned, we took the 10 most similar words to each bridging anaphor/anchor, in order to determine whether the word is in the list of comparable words to bridging anaphors/anchors according to Word2vec data.
2. *TF-IDF of word*—TF-IDF measure shows how important a word is to a document in a collection or corpus. This feature highlights the tendency anchors have to being in given information.
3. *Linear Distance*—the distance from the beginning of the text to the anchor, calculated in words.
4. *Lemma*—is it a lemma match to one of the anchors lemmas annotated in first part of the corpus?
5. *Type of syntactic link from the NP's head to word*—The MaltParser syntactic link type from the head of this word to the word. MaltParser uses a set of syntactic relations developed for SyntagRus [Boguslavsky et al 2006].
6. *Case*—a case automatically tagged by FreeLing.
7. *Syntax distance*—the shortest way from the bridging anaphor to the sentence root in the dependency tree is automatically built by MaltParser.
8. *Animacy*—an animacy or inanimacy automatically tagged by FreeLing.

3.2.2. Step 1. Feachors for Bridging Anaphors Prediction

For bridging anaphor recognition, we used the same number of features as the anchor feature set. The Classifier tends to identify nearly 70% of all nouns in the text as potential bridging anaphors. This is one of the reasons that there are no articles in Russian. On the other hand, alternative markers of definite NPs, such as deictic and possessive pronouns, seem to have a narrower distribution in Russian than in Romanic and Germanic languages where definite NPs are usually considered typical bridging anaphors. The fact that anaphors are generally less specific than antecedents, this can reflect on bridging anaphor recognition. These considerations need further exploration.

3.2.3. Step 1. Feachors for Bridging Pairs Prediction

The features used to train the classifier for bridging-pairs recognition include:

1. *Linear Distance*—linear distance between the bridging anaphor and the anchor.
2. *Probability of anchor*—estimated probability of the potential anchor computed by our first step classifier.

3. *Probability of bridging anaphor*—the estimated probability of the potential bridging anaphor computed by our first step classifier.
4. *Lemma of bridging anaphor*—is it a lemma match to one of the bridging anaphors lemmas annotated in first part of the corpus?
5. *Syntactic distance*—the shortest way from the bridging anaphor to the sentence root plus the shortest way from the anchor to the sentence root; if the bridging anaphor and anchor are in different sentences we just add 2, because we consider the texts a main root.
6. *Lemma of anchor*—is it a lemma match to one of the anchors lemmas annotated in first part of the corpus?
7. *Case of Bridging anaphor*—a case of the potential bridging anaphor automatically tagged by FreeLing
8. *Case of anchor*—a case of the potential anchor automatically tagged by FreeLing.

4. Experimental Results

For all experiments, we used cross-validation techniques for training, and an AUC measure for evaluating results. AUC is a square under the Receiver Operating Characteristic (ROC) curve. The ROC curve shows a correlation between the true positive rate (TPR) and the false positive rate (FPR) as seen in the graphs in section 4.3. An advantage of this measure is discussed in [Ling 2003]. AUC measure is a common measure for machine-learning experiments and classifier evaluation. The F-measure was also determined so that we could compare our results to related studies.

4.1. Machine-learning Algorithms

At the start of the study, we applied different machine learning algorithms to train the classifier. Three algorithms that are considered to be the least sensitive to correlated features that are common in natural language data are: Random Forest, Logistic Regression, and Decision Tree. For this study we chose Random Forest, which produced the most reliable results. For each algorithm, we tried different options, but such technical details are beyond the scope of this paper. The AUC measure for predicting anchors, bridging anaphors, and bridging pairs are given in Table 1.

Table 1. The application of different machine-learning algorithms results

	Random Forest	**Logistic Regression**	**Decision Tree**
AUC—Anchors	0.981	0.94	0.85
AUC—Bridging anaphors	0.969	0.93	0.92
AUC—Bridging pairs	0.92	0.79	0.7

4.2. Feature Importance

The semantic similarity feature is the most important feature of our set, followed by TF-IDF. Other features are less important, but all the features working in conjunction provide significant improvement.

Table 2. Anchor's features and features' importance

Anchor's Feature	Feature Importance	AUC without Feature
Semantic similarity to anchor/bridging anaphor	0.54	0.85
TF-IDF of word	0.17	0.98
Distance in words from the beginning of the text	0.07	0.98
Lemma	0.06	0.97
Type of syntactic link from the NP's head to the word	0.03	0.97
Case	0.02	0.98
Syntax distance from word to root of sentence	0.02	0.98
Animacy	0.01	0.97

Table 3. Bridging pairs' features and feature contributions

Bridging Pair's Feature	Feature Importance	AUC without Feature
Distance from bridging anaphor to anchor in words	0.25	0.82
Probability of anchor	0.23	0.83
Probability of bridging-anaphor	0.19	0.84
Lemma of bridging-anaphor	0.09	0.85
Syntactic distance from bridging anaphor to anchor	0.08	0.85
Lemma of anchor	0.05	0.88
Case of bridging anaphor	0.04	0.87
Case of anchor	0.03	0.88

4.3. Bridging Resolution Results

Our trained classifier shows strong results in anchor recognition (AUC=0.981) and weaker results for bridging anaphor recognition (AUC = 0.969). Poor bridging anaphor recognition impacts bridging pairs recognition in turn and for bridging pairs AUC=0.92. This result is not as high as we hoped it would be, but it is a satisfactory preliminary result and we believe it can be improved by extending the corpus and optimizing feature sets.

All the results obtained are presented in the charts below (Fig. 1–3). Included; are (ROC) curves. The ROC curve shows a correlation between the true positive rate (TPR) and the false positive rate (FPR). The AUC measure is the Area Under the Curve. For result "by chance" AUC=0.5

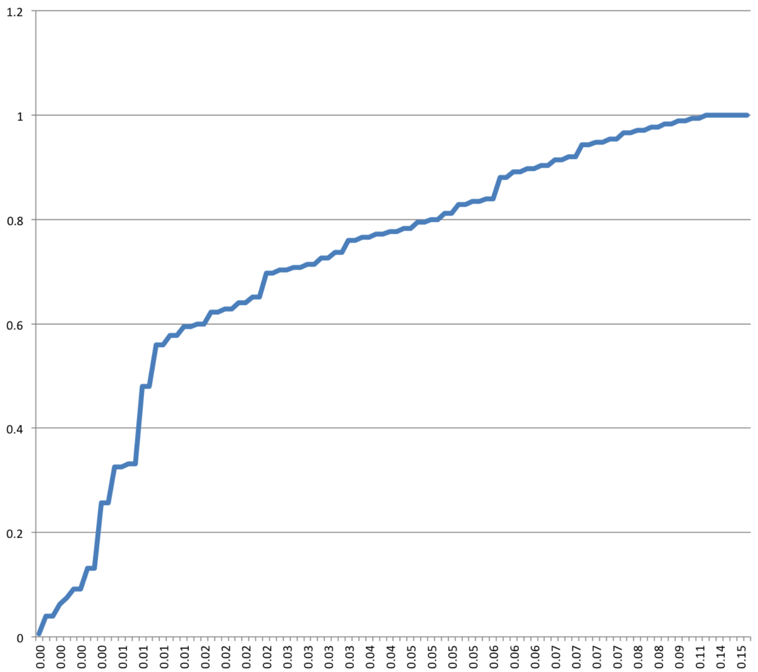

Fig. 1. ROC for Anchors. TPR is vertical, FPR is horizontal, AUC = 0.981

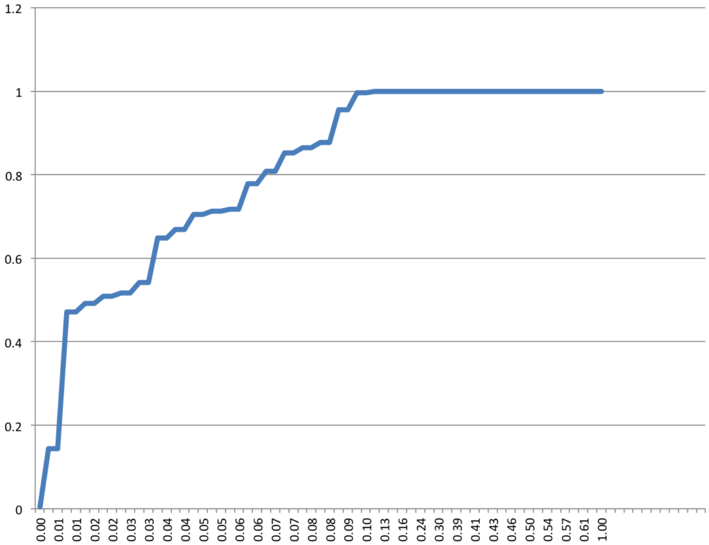

Figure 2. ROC for Bridging Anaphors. TPR is vertical, FPR is horizontal. AUC = 0.969

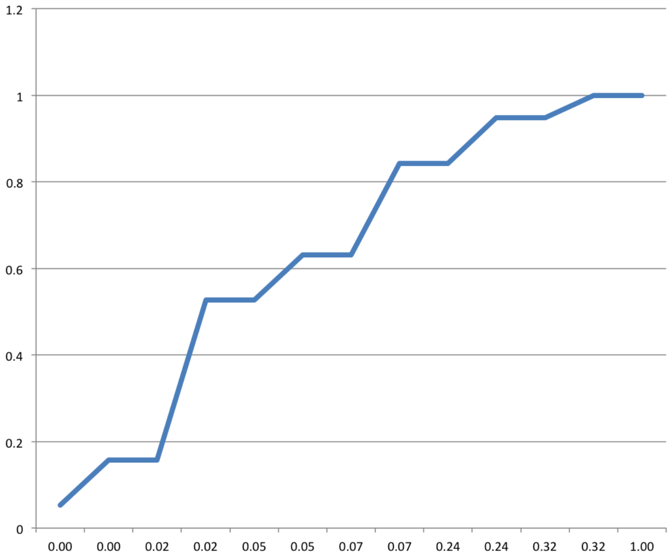

Figure 3. ROC for Bridging Pairs. TPR is vertical, FPR is horizontal. AUC = 0.92

F-measure was determined for bridging-anaphor and anchor recognition and bridging resolution. We can vary precision, recall and F-measure values by assigning the value of confidence score—the score shows the level of system confidence of the bridging element, anchor or pair. For bridging elements and anchors we vary the value of the rate in order to maximize precision. For bridging pairs we try to maximize F-measure. Maximum F-measure corresponds to a confidence rate of 0.35.

Results are presented in the tables bellow:

Table 4. Precision, recall and F-measure for bridging elements and anchors

	Bridging Element	**Anchors**
Precision	1	0.98
Recall	0.21	0.20
F-measure	**0.35**	**0.61**

The Table below shows the results for our bridging resolution system:

Table 5. Precision, recall and F-measure for bridging pairs

Precision	0.58
Recall	0.73
F-measure	**0.65**

4.4. Dependency Between Corpus Size and Bridging Recognition Quality

In Figures 4–5, the correlation between corpus size and classifier results quality is shown. The red line is the AUC computed for 100% of our data, 90% of our data, 80% and so on. Vertical intervals are the AUC measure dispersion between different runs of the classifier, which were trained on the mentioned corpus size. We have provided 10 runs for each variant of corpus size, from 100% to 40%. We changed the corpus size with 10% intervals.

The growth of the curves while using almost 60% of the data set for the anchor chart and close to 50% of the data set for bridging pairs was not the typical growth usually seen. The function should decrease monotonically from 100% to 40–50%. These growths could be due to a variety of specific text in training data.

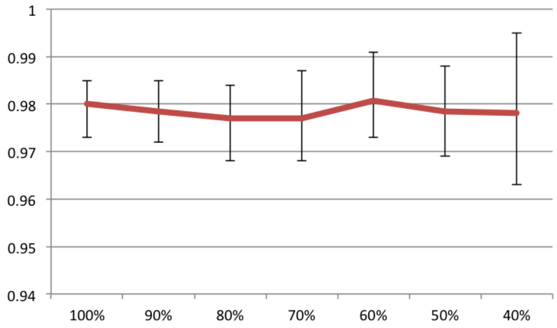

Figure 4. Dependency between corpus size and anchor recognition

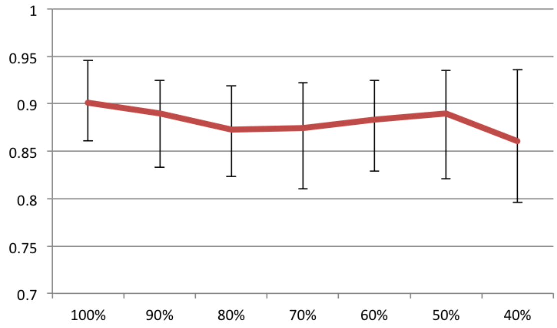

Figure 5. Dependency between corpus size and bridging-pairs recognition

5. Discussion

Automatic bridging resolution and recognition is still in its early stages of development. All other projects use different approaches to the sets of restrictions to bridging, so it’s difficult to compare our results with other bridging resolution and recognition studies. We found that our results (F-measure = 0.65) are high enough for such a complicated task, when compared to the F-measure used for the bridging resolution system for German [Klenner, Manfred, et al.], which is varied between 0.58 to 0.61.

Despite the preliminary results being adequate, we are going to continue our work with the goal of improving the level of automatic bridging resolution. Firstly, we plan to extend our corpus and optimize a feature set, extending the corpus should increase the result of bridging resolution. As shown in the extension in Figures 4–5, we have not yet reached a plateau, where increasing the data does not greatly influence the results. In relation to the features, we want to first improve the quality of syntactic features. For instance, currently we use all syntactic link types provided by MaltParces to compute the "type of syntactic link" feature while training the classifier; MaltParser distinguished more than 60 types of syntactic links. It is apparent that dividing all these syntactic link types into several groups so that the feature will have ten times less values results in a more effective feature. Also, we want to add a feature: "in one sentence" for bridging pairs, we expect that this will balance the "syntactic distance" feature, which is better than just adding 2, in the case of the bridging anaphor and anchor being in different sentences. There are also other features that are currently being considered for implementation.

Our approach for bridging resolution is simple. There is no need to use complicated, pre-prepared resources such as WordNet; these are only accessible for a small number of languages. To train a classifier, one simply needs a small, manually-annotated, corpus and automatic-annotation tools, which are well developed for a multitude of languages. Therefore, we hypothesize that our method can be applied to different types of bridging and different languages.

This work was supported by grants 16-04-01640 A and 15-07-09306 from RFBR.

References

1. *Boguslavsky, I., Grigorieva, S., Grigoriev, N., Kreidlin, L., & Frid, N.* (2000, July). Dependency treebank for Russian: Concept, tools, types of information. In Proceedings of the 18th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, pp. 987–991
2. *Clark, H. H.* (1975, June). Bridging. In Proceedings of the 1975 workshop on Theoretical issues in natural language processing, Association for Computational Linguistics, pp. 169–174
3. *Fan, J., Barker, K., & Porter, B.* (2005, October). Indirect anaphora resolution as semantic path search. In Proceedings of the 3rd international conference on Knowledge capture ACM, pp. 153–160
4. *Hou, Y., Markert, K., & Strube, M.* (2013). Global Inference for Bridging Anaphora Resolution. In HLT-NAACL pp. 907–917
5. *Klenner, M., Tuggener, D., Fahrni, A., & Sennrich, R.* (2010, August). Anaphora resolution with real preprocessing. In International Conference on Natural Language Processing, Springer Berlin Heidelberg pp. 215–225
6. *Lassalle, E., & Denis, P.* (2011, October). Leveraging different meronym discovery methods for bridging resolution in French. In Discourse Anaphora and Anaphor Resolution Colloquium, Springer Berlin Heidelberg, pp. 35–46

7. *Ling, C. X., Huang, J., & Zhang, H.* (2003, June). AUC: a better measure than accuracy in comparing learning algorithms. In Conference of the Canadian Society for Computational Studies of Intelligence (pp. 329–341). Springer Berlin Heidelberg.
8. *Lüngen, H.* (2008). RRSet-Taxonomy of rhetorical relations in SemDok. Interne Reports der DFG-Forschergruppe, 437
9. *Nivre, J., Hall, J., & Nilsson, J.* (2006, May). Maltparser: A data-driven parser-generator for dependency parsing. In Proceedings of LREC , Vol. 6, pp. 2216–2219
10. *Poesio, M., Delmonte, R., Bristot, A., Chiran, L., & Tonelli, S.* (2004-A). The VENEX corpus of anaphora and deixis in spoken and written Italian. University of Essex.
11. *Poesio, M., Mehta, R., Maroudas, A., & Hitzeman, J.* (2004-B, July). Learning to resolve bridging references. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, p. 143
12. *Poláková, L., Mírovský, J., Nedoluzhko, A., Jínová, P., Zikánová, S., & Hajicová, E.* (2013). Introducing the Prague Discourse Treebank 1.0. In IJCNLP (pp. 91–99).
13. *Roitberg, A., & Nedoluzhko, A.* (2016). Bridging Corpus for Russian in comparison with Czech. Coreference Resolution beyond OntoNotes, p. 59.
14. *Sasano, R., & Kurohashi, S.* (2009, August). A probabilistic model for associative anaphora resolution. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3, Association for Computational Linguistics, pp. 1455–1464