Improving Answer Extraction For Bangali Q/A System Using Anaphora-Cataphora Resolution

Conference Paper · December 2018

DOI: 10.1109/CIET.2018.8660888

3 authors:

Shomi Khan
Shahjalal University of Science and Technology
2 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Khadiza Tul Kubra
Shahjalal University of Science and Technology
1 PUBLICATION **4** CITATIONS
SEE PROFILE

Md Mahadi Hasan Nahid
Shahjalal University of Science and Technology
16 PUBLICATIONS **43** CITATIONS
SEE PROFILE

International Conference on Innovation in Engineering and Technology (ICIET) 27-29 December, 2018

Improving Answer Extraction For Bangali Q/A System Using Anaphora-Cataphora Resolution

Shomi Khan
Department of Electrical & Electronics Engineering
Shahjalal University of Science & Technology
nkskl6@gmail.com

Khadiza Tul Kubra
Department of Mathematics
Shahjalal University of Science & Technology
ktk.sust2015@gmail.com

Md Mahadi Hasan Nahid
Department of Computer Science & Engineering
Shahjalal University of Science & Technology
nahid-cse@sust.edu

***Abstract*—Human Computer Interaction (HCI) is a field of study to interact between humans (user) and computers on the design of computer technology. Question Answering (QA) system is one of the parts of HCI and a process of Information Retrieval (IR) in Natural Language Processing (NLP). In this research, it is attempted for a Bangla Question Answering System with simple sentences and experimented the system for both Bangla and English language. And it is tried to perform with semantic and syntactical analysis. Furthermore, for Bangla, a word net is constructed to demonstrate the system process. Our proposed method is a model in which it is easy for users to get most possible exact answer to their question easily and reduces the complexity of using noun instead of pronoun for the requested answer with respect to the given question queries for Bangla. It improves better answer extraction than naive approach.**

Keywords—Question Answering (QA), Bangali Question Answering (QA), Information Retrieval (IR), Natural Language Processing (NLP), Semantic and Syntactical Analysis, Word Net, Human Computer Interaction (HCI), Anaphora, Cataphora.

I. INTRODUCTION:

This QA system now-a-days has become very demandable, smarter and challenging system. Users usually require for quick response with exact answers to their queries [1]. But accessing the exact answer to the respective queries from the web document is not an easy task [2]. And if the query is in Bangla, it becomes more challenging since there are few works in Bangla.

QA system provides short answer to a natural language query for the corresponding question using either a pre-structured database or a collection of natural language web documents and presents only the requested information [1]. Natural language processing (NLP) with Information Retrieval (IR) is used by most of the QA system to search required questions [2]. NLP provides the computer understanding and manipulation of human language. It stands for the interaction between human and computer.

In this research, it is demonstrated in a web document hierarchy and a word net for the exact answer prediction by semantic matching with anaphora-cataphora resolution. Word net is referred as a lexical database which contains words with their synonyms and relational words, noun, adjective, verbs with grouping [2].

There are few works done for Bangla QA system. In Bangla, there are some additional complexity to extract answer from the document. Sometimes it displays pronoun as the answer of the question, that leads the accuracy of the answer to less.

In this research, it is experimented to reduce the complexity to extract exact answers from the dataset and replace the pronoun by the most suitable noun using word net.

II. RELATED WORKS:

There are lots of research works on English Question answering System, Question classification, Taxonomies and so on. But there is very few works on Bangla.

Banerjee and Bandyopadhyay, 2012 has done a work on Bangla Question classification. They have studied work suitable lexical, syntactic and semantic features and Bengali interrogatives and has proposed single-layer taxonomy of nine coarse-grained classes and has achieved 87.63%accuracy of question classification in their work [3].

There are two approaches to classify questions,

i. Rule based approach [3] [4]
ii. Machine learning based approach [3] [4]

Some researchers use some hybrid approaches by to combine the two approaches [4].

This combine approaches have never been used for Bangla Question Classification by any researchers. Li and Roth, 2004 and Lee et al., 2005 have proposed 50 and 62 fine grained classes for English and Chinese QC [3] [4]. Lexical, syntactical and semantic features (Loni, 2011) are the three categories of the features in QC [3]. A question in the QC task similar to document representation in vector space model is represented by Loni *et al.* (2011), i.e., a vector which is described by the words inside the question [4].

Thus, Q (question) can be represented as:

$$\Theta = (\Omega_1, \Omega_2, \Omega_3, ..., \Omega_{N-}, \Omega_N) \quad (1)$$

Where, Θ= frequency of term $k$in question Ω, and N= total number of Terms [4].

978-1-5386-5229-9/18/$31.00©2018IEEE

Question taxonomies is the question categories set. Single-layer taxonomy for Bengali question type have been used by Banerjee and Bandyopadhyay, 2012 which has eight coarse-grained classes with no fine-grained classes[3] [4]. No other researches have been contributed so far for Bengali taxonomies [3].

Banerjee and Bandyopadhyay, 2012 have used three features, lexical, semantical and syntactical feature in QC Lexical features(f_{Lex}) are of wh-word, wh-word positions, wh-type, question length, end marker, word shape, Syntactical features(f_{Syn}) are of POS tags, head word and Semantic features(f_{Sem}) are of related words, named entity [4].

III. Dataset

To demonstrate our system, at first, a pair of simple affirmative sentences have been taken in which one contains a bag of noun and another pronoun corresponding to the noun of the first sentence. We have considered 50+ pair of sentences as dataset.

IV. PROPOSED ALGORITHM:

In our developed QA system, it is capable of finding exact answer of given questions from the document. It tokenize the question into words, pop the wh-type words. Then it finds the best match through the document for the question words. The flowchart of our system is shown in below:

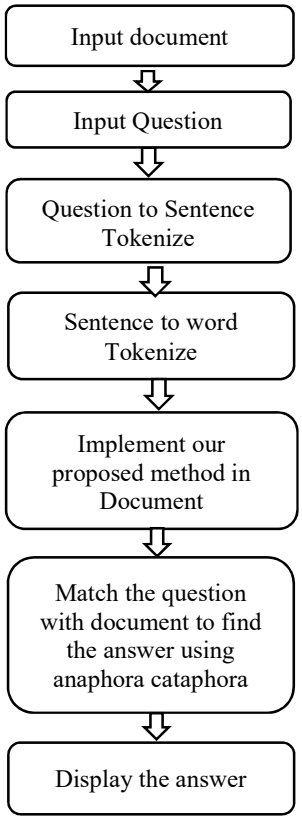

Fig. 1. Flow chart of our proposed QA system

In this system, a big problem was faced. For example-
1st sentence: কাঁঠাল খেতে ভারী মজা।
2nd sentence: এটি আমাদের জাতীয় ফল।
Question: আমাদের জাতীয় ফল কোনটি?
According to our system, answer would be— "এটি আমাদের জাতীয় ফল" which is not expected. It is needed to replace "এটি" with "কাঁঠাল". This is why a system is developed which can do it. Replacing the pronoun with its appropriate noun can be done by applying syntactic rules, semantic rules and reasoning type analysis.
A system has been developed to do this which is a combination of syntactic and syntactic rules. At first, we'll apply semantical rules. When it will fail to proceed we'll apply syntactical rule then.

i. Semantic rules:

To reach our goal, it is needed to know the context or topic of that sentence. For this reason, some semantic rules are applied here to get the appropriate noun from first sentence. To get the context we used some tag word of "বিশেষ্য", "বিশেষণ", "সর্বনাম" and "ক্রিয়াপদ" Bangla parts of speech.

For example,

The tag of "শাবিপ্রবি" is "বিশ্ববিদ্যালয়"
The tag of "বিশ্ববিদ্যালয়" is "বিশ্ববিদ্যালয়"
The tag of "পাবলিক বিশ্ববিদ্যালয়" is "বিশ্ববিদ্যালয়"
The tag of "সিলেট" is "বিভাগ"
The tag of "বিভাগ" is "বিভাগ"

Now,
1st sentence: শাবিপ্রবি সিলেট বিভাগে অবস্থানরত একটি পাবলিক বিশ্ববিদ্যালয়।
2nd sentence: এখানে EEE বিভাগ আছে।

Now it can be seen that, three words of first sentence have "বিশ্ববিদ্যালয়" tag and two words have "বিভাগ" tag. So, dominating tag is "বিশ্ববিদ্যালয়". So, the context of the first sentence is "বিশ্ববিদ্যালয়". Among the words of "বিশ্ববিদ্যালয়" tag only "শাবিপ্রবি" is the proper noun. So, "এখানে" will be replaced by "শাবিপ্রবি".

Now, another example is,

The tag of "পাইথন" is "প্রোগ্রামিং"
The tag of "প্রোগ্রামিং" is "প্রোগ্রামিং"
The tag of "প্রোগ্রামিং ভাষা" is "প্রোগ্রামিং"
1st sentence: পাইথন একটি প্রোগ্রামিং ভাষা।
2nd sentence: এটি মেশিন লার্নিং এর জন্য অনেক জনপ্রিয়।

From this example, it can be seen that three words have "প্রোগ্রামিং" tag and no other tags are found here. So, the context of the first sentence is "প্রোগ্রামিং". "এটি" will be replaced by "পাইথন".

Now,
The tag of "ঈশ্বরদী" is "থানা"
The tag of "থানা" is "থানা"
The tag of "পাবনা" is "জেলা"
The tag of "জেলা" is "জেলা"
1st sentence: ঈশ্বরদী হল পাবনা জেলার অন্তর্ভুক্ত একটি থানা।
2nd sentence: এখানে আমার দাদাবাড়ী।
Here, it can be seen that, there is 2 word that contains tag "জেলা" and other 2 words contains tag "থানা". Now it will be done a priority scoring by applying syntactic rules.

TABLE I. "জেলার" TAG:

Words	Score	Reason
পাবনা	2	"পাবনা" is a single word, so upper priority.
জেলার	1	"জেলার" is a word with the combination of two words "জেলা" and "এর". As it contains "এর, it gets lower priority. Details can be found in Syntactical rules.
Total	3	

TABLE II. "থানা" TAG:

Words	Score	Reason
ঈশ্বরদী	2	Single word
থানা	2	Single word
Total	4	

We can see that, the total score or priority of "জেলার" tag is 3 and "থানা" tag is 4. So, we'll consider the upper priority. Here, the tag of "থানা" is the upper priority and its word tag is "ঈশ্বরদী". So, we'll take "ঈশ্বরদী" as our most possible result.

The tag of "মিষ্টিকুমড়া" is "খাবার".
The tag of "ভিটামিন-'এ'" is "পুষ্টিগুণ".
The tag of "খাবার" is "খাবার"

1st sentence: মিষ্টিকুমড়া একটি ভিটামিন-'এ' সমৃদ্ধ।
2nd sentence: এটি চোখের জন্য খুব উপকারি।

Here, "মিষ্টিকুমড়া" and "ভিটামিন-'এ'" are noun in the first sentence. They both are single words. In the second word, "এটি" tag denotes for both "মিষ্টিকুমড়া" and "ভিতামিন-'এ'", since their priorities are equal.

From another example, we can see,

The tag of "মিষ্টিকুমড়া"is "খাবার".
The tag of "ভিতামিন-'এ'" is "পুষ্টিগুণ".
The tag of "খাবার"is "খাবার".
The tag of "সুস্বাদু" is "খাবার".

1st sentence: মিষ্টিকুমড়ায় ভিটামিন 'এ' আছে।
2nd sentence: এটি সুস্বাদু।

Till now, it can be seen that, there is no tag in second sentence. But in this example, we can find "সুস্বাদু" tag in the second sentence.
It can be seen that, there are two words with "খাবার" tag and one is "পুষ্টিগুণ" tag. In the second sentence we can find "এটি". It will give the direction to the "খাবার" tag. So, we'll consider "মিষ্টিকুমড়া" tag as expected answer, since in the first sentence there is "খাবার" only tag is "মিষ্টিকুমড়া".

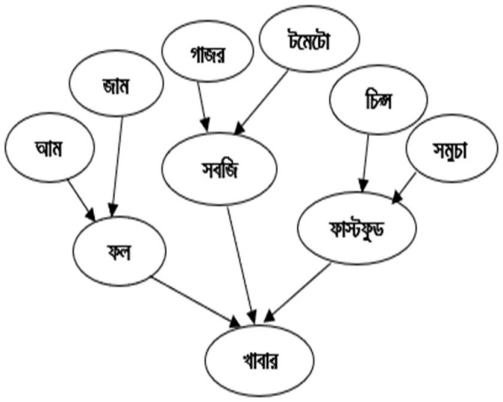

Fig. 2. Tag tree

This is the tag tree which is used in wordnet. Leaf nodes are proper noun. Internal nodes are their tag word. The internal are itself their own tag word. We will go upper tag until we don`t find max tag words.

Now, the total overview of the semantic rule is given below as a flow chart:

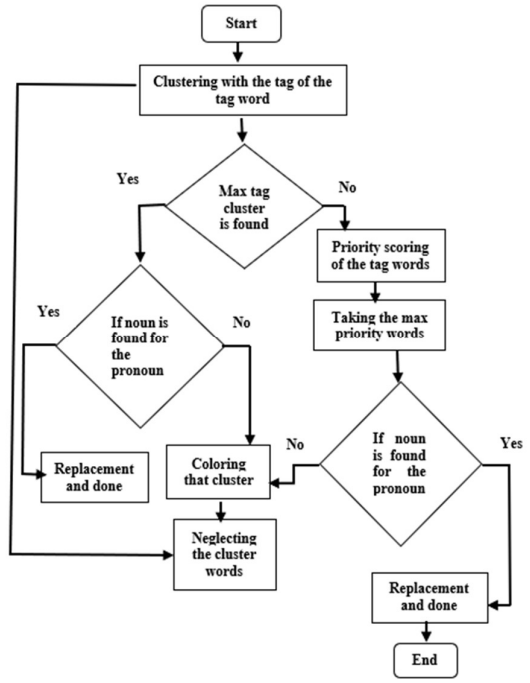

Fig. 2. Flow chart of the overview of the semantic rule

ii. Syntactical Rules:

In our system, the pronoun is replaced with its corresponding noun.
If only noun in the first sentence is got, then it will be directly replaced with the pronoun of 2nd sentence. For example,
1st sentence: রাতুল ভালো ছেলে।

2ndsentence: সে প্রতিদিন স্কুলে যায়।

Here, "রাতুল" is the only subject of first sentence. So, "সে" from the second sentence will be replaced by "রাতুল" from the first sentence.
But, sometimes there is more than one nouns.
For example,
1st sentence: রাহুল মিতার গান শুনে মুগ্ধ হয়ে গেল।

2ndsentence: সে খুশীতে হাততালি দিল।

Here, "রাতুল" is the noun who is acting and "মিতা" is the noun with "এর" and related to noun "রাহুল". This is why, we replaced "সে" with "রাহুল".
Another example,
For example,
1st sentence: শিক্ষক ছাত্রদের কাছে গেল।

2ndsentence: তারা তাকে দাড়িয়ে সালাম দিল।

Here, "তারা" is the pronoun which is in plural form. So, indicates "ছাত্রদের". "তাকে" is the pronoun which is in singular form. So, it indicates "শিক্ষক".
For example,
1st sentence: রাহুল মৃদুলের বাসায় গেল।

2nd sentence: সে তার জন্য একটা উপহারও নিয়ে গেল।

Here, "রাহুল" and "মৃদুল" are two nouns from the first sentence and "সে" and "তার" are pronouns from second sentence. There is a "এর" with "মৃদুল". So, in this sentence, "মৃদুল" is somehow connected with "রাহুল". "রাহুল" is the noun who is taking action. This is why it is taken "মৃদুল" as lower priority noun and replaced it with "তার" of second sentence. "তার" is the lower priority pronoun than "সে" for the same reason. "সে" from the second sentence is replaced by "রাহুল". Similarly, if any noun with "এর", "কে", "তে" etc. it was taken as lower priority noun and replaced it with lower priority pronoun "তার", "তাকে" etc.
1st sentence: রাহাত একটি খেলনা কিনে আনল।

2nd sentence: সেটি খুব সুন্দর।
Here, "সেটি" is an object type pronoun from the second sentence. So, it will search an object type noun from the second sentence. "রাহাত" and "খেলনা" both are noun in the first sentence. Here "রাহাত" is a subject type noun and "খেলনা" is an object type noun. So "সেটি" will replace "খেলনা".
Again,
1st sentence: যশোরের দই খেতে ভালো।

2nd sentence: এখানে এটি বিখ্যাত।

There are two nouns in the first sentence and two pronouns in the second sentence. Since, the only place type noun is "যশোর" and object type noun is "দই", the place type pronoun "এখানে" of the second sentence will be replaced by the noun "যশোর" and the object type pronoun. "এটি" will be replaced by the noun "সেটি".

V. RESULT ANALYSIS:

It has been taken fifty documents with ten question queries for each document in this study. But here is shown only ten examples in the table as sample.

TABLE III. RESULT SCROING TABLE

No	Document	Question	Naïve Approach	Our Proposed System	Score
1.	কুষ্টিয়ায় ভালো তিলের খাজা পাওয়া যায়। এখানে এটি খুব বিখ্যাত।	কোনটি কুষ্টিয়ায় খুব বিখ্যাত ?	এখানে এটি খুব বিখ্যাত	কুষ্টিয়ায় তিলের খাজা খুব বিখ্যাত।	(1/1)
2.	মিষ্টিকুমড়ায় ভিটামিন-এ আছে। এটি খেতে ভালো।	কোনটি খেতে ভালো?	এটি খেতে ভালো	মিষ্টিকুমড়া খেতে ভালো।	(1/1)
3.	পাইথন একটি	কোনটি খুব	এটি খুব জনপ্রিয়	পাইথন খুব জনপ্রিয়।	(1/1)

	প্রোগ্রামিং ভাষা। এটি খুব জনপ্রিয়।	জনপ্রিয় ভাষা?			
4.	রাহাত তাহসিনকে একটি উপহার দিল। এজন্য সে তাকে ধন্যবাদ দিল।	কে ধন্যবাদ দিল?	এজন্য সে তাকে ধন্যবাদ দিল।	এজন্য সে তাকে ধন্যবাদ দিল।	(0/2)
5.	রাজশাহীর আম খেতে খুব মজা। এটি এখানে ভালো ফলে।	আম কোথায় ভালো ফলে?	এটি এখানে ভালো ফলে।	আম রাজশাহীতে ভালো ফলে।	(2/2)
6.	শাওন দরজায় এসে দাঁড়ালো। মনি তাকে ভেতরে আসতে বলল।	কে কাকে ভেতরে আসতে বলল?	মনি তাকে ভেতরে আসতে বলল।	ফারহান শাওনকে ভেতরে আসতে বলল।	(2/2)
7.	বাপ্পা ত্রিকোণমিতির চেয়ে জ্যামিতি ভালো পারে। কারণ, সে এটা পছন্দ করে।	বাপ্পা কোনটা পছন্দ করে?	কারণ, সে এটা পছন্দ করে।	কারণ, বাপ্পা জ্যামিতি পছন্দ করে।	(1/1)
8.	রাতুল সাকেরকে একটি উপহার দিল। সে সেটি গ্রহন করল।	কে উপহার গ্রহন করল?	সে সেটি গ্রহন করল।	রাতুল উপহার গ্রহন করল।	(0/2)
9.	রাখি মেলায় গিয়ে একটি বাঁশি কিনল। সে সেটি ভেঙে ফেললো।	সে কি ভেঙে ফেলল?	সে সেটি ভেঙে ফেললো।	রাখি খেনলাটি ভেঙে ফেললো।	(1/1)
10.	আয়েশার সাথে অনামিকা ঝগড়া করেছে। তারা দুইজন দুইজন কে অপছন্দ করে।	কে কাকে অপছন্দ করে?	তারা দুইজন দুইজন কে অপছন্দ করে।	তারা দুইজন দুইজন কে অপছন্দ করে।	(0/2)

The scoring has been demonstrated in the table. Each problem has been scored due to the number of pronouns found according to the required number. Three problems have been demonstrated, that fails to find the required nouns and seven problems that finds the nouns and also extract the single words for exact answer. When it fails, it is scored 0 out of 1. When there are nouns for the pronoun it is scored 1 out of 1.

$$Accuracy=\frac{No.of\ correct\ answer}{Total\ no.of\ of testing\ sample} \quad (2)$$

$$=\frac{1+1+1+0+2+2+1+0+1+0}{15}\times 100\%$$

$$=60\%$$

TABLE IV. ACCRACY MEASURING TABLE

No of observations	Accuracy given by naïve approach	Accuracy given by our proposed system
1.	60%	100%
2.	50%	60%
3.	60%	60%
4.	80%	100%
5.	60%	70%
6.	50%	60%
7.	60%	70%
8.	60%	60%
9.	70%	80%
10.	60%	80%
Total in average	**61%**	**74%**

Here, in observation 4, 8, 10 are based on reasoning fact. There are many types of reasoning facts like, induction, deduction, counting etc. Dynamic Memory Allocation (DMA) can do this type of processing. It also needs some basic knowledge. Just like,

1) If someone do something bad, he will feel guilt later.
2) If a person thanks someone, he will welcome him.
3) Fast food is dangerous for health. Etc.

Semantic Memory (SM) can be used to save and use the basic knowledge or static knowledge.

Actually, to improve our accuracy it is needed to use the combination of both DMA and SM. It wil be done by us. But in our previous system, it fails to show satisfactory results. Which is comparatively immobile than our developed system.

VI. CONCLUSION:

In this research a usable QA system have been implemented. During this implementation, some challenges are faced, for example, our system only works for simple sentences. It will be puzzled and give incorrect

answer if the sentence is very complex. The Bangla word dataset is not so much enriched. And for Bangla, the contribution is so little. It needs a lot of time and effort. In future, it can be possible ensuring a better result by using better methods, techniques and resources.

In this system 60% accuracy have been found. A better accuracy can be found by further processing in future.

References

[1] Lende, Sweta P., and M. M. Raghuwanshi. "Question answering system on education acts using NLP techniques." *Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), World Conference on*. IEEE, 2016.

[2] Jayalakshmi, S., and Ananthi Sheshasaayee. "Automated question answering system using ontology and semantic role." *Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on*. IEEE, 2017.

[3] Banerjee, Somnath, and Sivaji Bandyopadhyay. "Ensemble approach for fine-grained question classification in bengali." *27th Pacific Asia Conference on Language, Information, and Computation*. 2013.

[4] Banerjee, Somnath, and Sivaji Bandyopadhyay. "An Empirical Study of Combing Multiple Models in Bengali Question Classification." *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. 2013.

[5] Iyyer, Mohit, et al. "A neural network for factoid question answering over paragraphs." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

[6] Yih, Wen-tau, et al. "Question answering using enhanced lexical semantic models." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2013.

[7] Andreas, Jacob, et al. "Learning to compose neural networks for question answering." *arXiv preprint arXiv:1601.01705*(2016).

[8] Cooper, Richard J., and Stefan M. Ruger. "A simple question answering system." *TREC*. 2000.

[9] Moldovan, Dan, et al. "Lasso: A tool for surfing the answer net." *TREC*. Vol. 8. 1999.

[10] Tellex, Stefanie, et al. "Quantitative evaluation of passage retrieval algorithms for question answering." *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003.

[11] Grishman, Ralph. "Information extraction: Techniques and challenges." *International Summer School on Information Extraction*. Springer, Berlin, Heidelberg, 1997.

[12] Hovy, Eduard, et al. "Question answering in webclopedia." *TREC*. Vol. 52. 2000.

[13] Senapati, Apurbalal, and Utpal Garain. "Guitar-based pronominal anaphora resolution in bengali." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vol. 2. 2013.

[14] Tazakka, Tazbeea, Md Asifuzzaman, and Sabir Ismail. "Anaphora Resolution in Bangla Language." *International Journal of Computer Applications* 154.9 (2016).

[15] Sanjay Chatterji, Arnab Dhar, Biswanath Barik, Sudeshna Sarkar and Anupam Basu, “Anaphora resolution for bengali, hindi, and tamil using random tree algorithm in weka.” In Proceedings of the ICON-2011, 2011.

[16] Sikdar, Utpal Kumar, Asif Ekbal, Sriparna Saha, Olga Uryupina, and Massimo Poesio. “Anaphora Resolution for Bengali: An Experiment with Domain Adaptation”, Computación y Sistemas 17, no. 2, 2013 : 137-146.

[17] Fedele, Emily, and Elsi Kaiser. "Looking back and looking forward: Anaphora and cataphora in Italian." *University of Pennsylvania Working Papers in Linguistics* 20.1 (2014).

[18] Bharadwaj, Rohit G., et al. "A Naïve Approach for Monolingual Question Answering." *CLEF (Working Notes)*. 2009.