IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

Named Entity Disambiguation Anaphora Resolution and Question Answering in Speech

C H. Chithira

Dept. of Computer Science, Government Engineering College Sreekrishnapuram, Palakkad, India

Corresponding Author:chithirach@gmail.com, Tel.: +917561896322

Available online at: www.ijcseonline.org

Accepted: 12/Dec/2018, Published: 31/Dec/2018

***Abstract*— Speech Recognition, Named Entity Disambiguation(NED), Anaphora Resolution and Question Answering(QA)** are some of the major areas of research in Natural Language Processing(NLP). Speech recognition systems will convert the human voice into corresponding text. Named Entity Disambiguation will identify the entity types in the given text. Anaphora Resolution is the process of finding antecedents of an anaphor and it has become a challenging task in the Computational Linguistics and Natural Language Processing(NLP). Question answering will develop systems that automatically answers the questions in natural language. Question answering system and Named Entity Disambiguation are the important fields intended to enhance the performance of the Information Retrieval System. This work mainly focused on integrating the above three tasks in a single system. Moreover this system will perform these tasks by receiving human speech as input and then converting it into the corresponding text. After the text conversion, Entity Disambiguation is performed on it. Real time speech input are given as the input. Anaphora resolution is also integrated with the disambiguation phase. The Question Answering System completely depends on the efficiency of entity disambiguation. The Answers are retrieved by searching based on these disambiguated entities. Anaphora Resolution will provide greater support for those entities that are not correctly identified in the disambiguation phase. Also the problem that anaphors that couldn't find proper antecedents can easily find the correct ones since their entity type is correctly identified. The Question Answering System will perform better if the entities are correctly identified and disambiguated.

***Keywords*—** Speech Recognition, Named Entity Disambiguation, Anaphora Resolution, Question Answering, Natural Language Processing(NLP), Computational Linguistics.

I. Introduction

People are very busy in the present world, they are so adjusted with this fast scenario. They are always interested in getting things instantly to save the time. Earlier mobile phones are replaced by touch screens now, for the ease of human effort. The keypads in the laptops reduce the use of keyboards, because we can operate it by just touching and moving our fingers instead of typing. Now they are replaced by the Speech recognition systems. We just have to ask anything that we want to know. The speech converter systems will convert it into the corresponding text. Google speech is one of the systems in which user can search information through the browser just by asking the query to the PC's. They have to type nothing for that. This work is based on such a system. The human speech is converted into its corresponding text. Speech recognition systems are also called as Automatic Speech Recognition systems, Computer Speech Recognition systems or Speech to Text (STT). This is an active area of research in the field of Computational Linguistics, Computer science and Electrical Engineering. Named Entity Disambiguation(NED) and Anaphora Resolution are performed on the converted text. Named Entity Disambiguation will decide the identity of entities in the text. It is also known as Named Entity Linking(NEL), Named Entity Normalization(NEN), Named Entity Recognization and Disambiguation(NERD) Named Entities are basically the atomic parts of a given text ie. Names of persons, Organizations, Locations, Expressions of time, Quantities, Monetary values, Percentages etc [1]. The Entity Disambiguation can enhance the performance of Information Retrieval systems. It is highly significant for the semantic search. Anaphora Resolution will find the antecedent of an anaphor. Anaphor is the reference pointing to the previously occured item. Antecedent is the entity that the anaphor refers. Anaphora resolution has become a challenge to the natural language processing and computational linguistics. Question answering systems can automatically answer the questions asked by the humans and the questions are in natural language [2]. The question answering systems will retrieve precise answers for the given user queries. It depends on the efficiency of a search corpus. Larger corpus size will led to better QA performance. This work aims to perform Named Entity

Disambiguation, Anaphora resolution and question answering in speech. Human speech are converted to the text. After disambiguation, anaphora resolution is performed on it. The question answering system depends on the correct disambiguation of entities. DBpedia Spotlight is used as the dataset for disambiguation and question answering. Postagging of the input sentences are done using Stanford postagger [3] .

Rest of the paper is organized as follows, Section I contains the introduction of the above work. Section II contain the related work. Section III contains the methodology used for obtaining the result. Section IV explains the achieved results and comparison with previous results and Section V concludes research work with future directions

II. Related Work

The work "**NEED4Tweet: A Twitterbot for Tweets Named Entity Extraction and Disambiguation**" by Mena B. Habib et al. [1] introduces a twitterbot named NEED4Tweet for named entity extraction and disambiguation in tweets. Twitter is an important social media that contains rapidly changing information generated by millions of users. Twitterbot is a program that is used to produce automated posts on twitter by receiving the tweet, processing it and sending a reply message contains a link to a page that shows the generated annotations. The proposed system consists of three phases : NE candidate generation, Disambiguation, NE candidate filtering. The output from the two candidate generation methods, Tweet segmentation and KB(knowledge base) are integrated to generate the candidates.

The work **"Named Entity Disambiguation using Linked Data"** by Danica Damljanovic et al. [2] mainly focused on integrating a state- of-the-art named entity tool with Linked Data-based similarity measures and proved thet their algorithm can improve disambiguation accuracy on a subset of Wikipedia user profiles. Their algorithm will identify named entities in text and each of them will be attached to the correct DBpedia URI (Uniform Resource Identifier). ANNIE extraction system from GATE is used for extracting Named Entities from the text. It will generally focus to produce named entity types such as Person, Organization, Location. Since it also resolves coreference, entities with the same meaning are also linked. Large Knowledge Gazetteer (LKB), which is the ontology based gazetteer of the GATE is used to link the correct URI with the entities. They compare both ANNIE and LKB in terms of efficiency of disambiguation. They introduce an algorithm by consolidating the output of ANNIE and LKB.

"A Technique for Anaphora Resolution of Text" by Vipin Kumar et al. [3] proposed an Anaphora Resolution method for the information management. This method will find referents of the verb phrase form and it also distinguishes between pleonastic 'it' and anaphoric 'it'. This method resolves the anaphora which is referred to after an interval of multiple sentence. These referents are stored in a list according to their occurance. Recency is used to select the correct referents, if the agreement features not suffices to estimate the correct referent for an anaphor. WordNet lexical database are used to compare the synonym of an anaphor with possible referents. The proposed system works well on plain text documents. They named the system as 'New Resolution System'. The important steps in the implementation of the system are Preprocessing, Anaphora Detection and Resolution. The main design constraints are Number agreement, Gender agreement, Person and Case agreement Syntactic constraints, Selectional constraints.

"**EM Works for Pronoun Anaphora Resolution**" by Eugene Charniak et al. [4] introduced an algorithm that uses expectation maximization in an unsupervised manner. Since it is using the Expectation Maximization approach, it mainly focus on resolving Personal Pronoun Anaphora. They argue that their system resolves personal pronouns like subjective, objective, possessive, reflexive. They explained about the two categories of personal pronoun, anaphoric and non anaphoric pronouns. Personal pronoun has three properties namely person, number and gender. The algorithm will first decide whether the pronoun is anaphoric P (anaphoric). For the anaphoric pronoun, it will find the possible antecedent. Current or two previous antecedents are considered here. Then select antecedent based upon a distribution P (anaphora|context). Then generate pronoun's person P(person|antecedent), person's gender P(gender| antecedent), person's number P (number|antecedent), governor/relation like P (governor/relation|antecedent) and also non-anaphoric it P (it|nonanaphoric), P (governor/relation|non− anaphoricit). Smoothing method is implemented using Kneser-Ney smoothing.

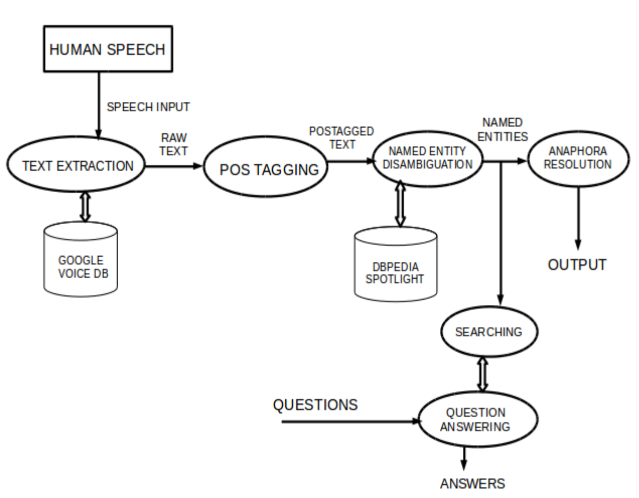
"Design of the Effective Question Answering System by Performing Question Analysis using the Classifier" by Gayatri Chavan et al. [5] proposed an open domain question answering system which is related to natural language interface to the database (NLIDB). This system takes natural language query input and gives appropriate answers from the manually created knowledge base (structured database). The important steps for the proposed system is the use of classifier for the identification of tables and columns in a structured database for the incoming questions and performing free text retrieval to lookup answer. Statistical classifier trained on data from TREC QA task is used here. The main advantage of this system is the aviodance of expensive text analysis. The knowledge base for this system consists of commonly occuring question types that are extracted and stored in structured database for lookup at question time. The important modules of the question answering system are the retrieval module and classifier module. The retrieval module consists of text representation and the retrieval query formulation.

The work “**Named Entity Disambiguation in a Question Answering System”** by Marcus Klang et al. [6] described how named entity disambiguation can be used to merge the entities in the question answering system. An integration of named entity disambiguator in swedish language to a question answering system named Hajen can be viewed here. The named entity disambiguation part is done by connecting a sequence of words or proper nouns to a unique identifier. Wikidata is used as an entity repository. Wikidata identifiers will gather properties from the Wikipedia online encyclopedia and the infobox tabulated information associated to some of its articles. Here the entity linking consists of three steps mainly part of speech tagging using Stagger, linking the string to possible wikidata identifiers and disambiguation using popularity, commonness and a boolean context method. This named entity linker is then integrated with baseline question answering system.

III. Methodology

[1] Speech Recognition, Named Entity Disambiguation, Anaphora Resolution And Question Answering

Figure 1. Overall architecture for speech recognition, named entity disambiguation, anaphora resolution and question answering



A. Speech Recognition

The input given is human voice and the output is its corresponding text. The voice from physical hardware is Captured. Human voice are given as signals. They are compared with the Google voice Database. If it is in the voice database, then text corresponding to the voice signals in the dataset are retrieved. After that they are Stored in a textbox. After storing all those words, each word is send to the Language Database. Final output is produced from Language Database. If it is an unknown word (eg.Names) that is not in the voice database: then they are send to the Language Database. Each letter is treated individually and reproduce the same word(letter by letter). Relevant details should be given including experimental design and the technique (s) used along with appropriate statistical methods used clearly along with the year of experimentation (field and laboratory).

B. Named Entity Disambiguation and Anaphora Resolution

Named Entity Disambiguation will decide the identity of entities in the text. It is also known as Named Entity Linking (NEL), Named Entity Normalization (NEN), Named Entity Recognization and Disambiguation (NERD). Named Entities are basically the atomic parts of a given text ie. Names of persons, Organizations, Locations, Expressions of time, Quantities, Monetary values, Percentages etc. The proposed system will works as follows: The raw text sentence extracted from the speech is undergone segmentation. Then it is postagged using StanfordPOStagger. POS Tagger will assign tags for each word such as verb, noun, pronoun, adjective etc. Therefore postagging will identify the names or nouns in the given sentences even if it is a person, location, organization etc. After that these names are compared with the dataset, to know whether it is a named entity or not.

The dataset used here is DBpedia Spotlight. DBpedia will gives the entity types and it will attach correct URI (uniform resource identifier) to each entity. It also gives different categories in which this entity belongs to. It also gives the correct and precise description about the entities. All the above mentioned facts can be observed in the output that is displayed. Here the disambiguation part will find the webarticles in which this entities are present and sort the entities those that are having higher support. Searching through the whole article is a tidious task. Here comes the significance of DBpedia. DBpedia Spotlight will give the annotated data from the wikipedia articles. DBpedia will provide the datas of wikidata and those datas are annotated. Wikidata couldnot provide annotated data. Consider the example: ’Steve Jobs is the founder of Apple’, here ’ Apple’ is the word that creates ambiguity ie. whether it is a named entity or not. The disambiguation part will first consider the word ’Steve Jobs’ and check the dataset to find whether it is a named entity or not. The DBpedia will gives the information that it is a person. Then the disambiguation part will find the articles in which Steve Jobs is present. Then it considers the word ’Apple’ and found that it is an ambiguous entity. Then it will also search the articles in which ’Apple’ is present and sort the articles in which both entities are found. Finally it will consider the event ’founded’ and again sort the articles based on this event and select the article having higher support. Then finalize the entity type of ’Apple’ ie. it is an organization. The disambiguated entities are stored in the database. They are then displayed from this database. MYSQL is used as the database. Anaphora resolution is also integrated with it.

Anaphora Resolution determines the antecedent of an anaphor. Anaphora is combined with NED after postagging. A list of anaphora and noun phrases are generated after postagging. The gender identification will also be done in the case of pronouns (ie.’He’for male, ’she’ for female). After postagging, nouns and pronouns are send to the gender list for gender identification. From the list of pronoun and noun phrases, all possible pairs of anaphora and antecedents are generated.

Each pair is then filtered through agreement filter which checks for compatibility of each pair on the basis of agreement(person, gender, number) features. These pairs are then send to the noun list. Personal pronoun filter is applied then (third person pronoun ie. He, She, They) Comparing the pronoun with the entity just before it and retrieve that entity along with its pronouns if it matches. The ouput sentences with referential pairs are then given to the entity disambiguation. The final output will be disambiguated entities and their anaphors. Since the anaphora resolution is based on finding the anaphors of named entities, this work mainly concentrates on pronominal anaphora resolution. It doesn’t consider the cases like pleonastic ’it’. The intersentential anaphora are implemented succesfully.

C. Question Answering System

Question answering systems will develop systems that automatically retrieve answers in the natural language.

These systems will retrieve answers by querying the knowledge base. It is an answer driven approach that retrieves short and precise answers for the queries asked by users. The questions may be phrase based, full sentence or keyword based.

Question answering system deals with variety of question types like fact, list, definition, How, Why, hypothetical, semantically constrained, and cross-lingual questions. Basically there are two types of question answering system open domain and closed domain. Closed domain deals with questions under a particular domain. This work mainly focused on the Factoid queries and Definition queries. The question answering system depends on NED. It works based on the same idea of the Named Entity Disambiguation. It will give exact answers only if the disambiguation part works correctly. For example: Consider the question ’where did Mahatma Gandhi born ?’, here the question word ’Where’ refers to a location.

It is already predefined that Who refers to a person, Where refers to a location, Which refers to a location or a name, What refers to a person or a location etc. So it is clear that the above example checks for a location. After that it considers the named entity ’Mahatma Gandhi’ and the event ’born’ and sorts the web articles showing the entity and the event.

Finally the location having highest support is chosen as the answer for the query. This question answering system also depends on DBpedia Spotlight. It is connected with the disambiguation phase and its output.

IV. Results and Discussion

For evaluating the work, two datasets named ’Politics’ and 'Cricket' are chosen. Both datasets contain 300 entities and in total there are 600 entities. Disambiguation with Anaphora and without Anaphora are done for these two domains.

Named Entity Disambiguation is done during the intermediate phase of the project ie. before integrating anaphora to it.

After performing Anaphora Resolution, the datasets are disambiguated with Anaphora. The datasets are created manually by taking data from the wikipedia articles. The first dataset named ’Politics’ consists of descriptions about the Indian political leaders and parties having 300 named entities and these articles are manually selected from the wikipedia articles to perform evaluation. The second dataset ’cricket’ also consists of articles about indian cricket and cricketers. It also consists of 300 entities. The evaluation measures chosen are: Precision, Recall and F-measure.

Table 1. Precision, Recall and F-measure of Named Entity Disambiguation and Anaphora Resolution

	Dataset 1: Politics			Dataset 2: Cricket		
	precision	Recall	F-measure	precision	Recall	F-measure
Disambiguation without Anaphora	0.85	0.82	0.83	0.84	0.85	0.84
Disambiguation with Anaphora	0.89	0.86	0.87	0.84	0.88	0.85

It will enable us to conclude that which type of entity it is. we can easily recognize those entities that are not identified in the disambiguation part. Not only the person, but also the other named entities. The anaphora resolution also identifies the pleonastic ’it’. The anaphora resolution will provide more support for the entities that are not recognized. Samething can be observed in the case of dataset 2(cricket). While comparing both datasets, dataset 1 has more precision, recall and f- measure than dataset 2. But the anaphora resolution does not identify the pronominal anaphors if we give sentences having multiple persons as antecedents. There occurs some sort of confusion in the case of pleonastic it. Sometimes the pleonastic it and the anaphor it that refers a non-animistic(non living thing) named entity more specifically an organization makes confusion.

Table 2. Precision, Recall and F-measure of Question Answering System

	Recall	Precision	F-measure
Without Disambiguation	0.77	0.76	0.76
With Disambiguation	0.78	0.77	0.77

The question answering system are found to be more efficient when compared to other systems. The evaluation of the question answering system is done manually. Questions from different domains are taken and given to the question answering system randomly. The efficiency of question answering system depends on the correct disambiguation of entities. The Factoid queries (Wh questions like what, which, where), Definition queries(Questions start with 'what' which needs small descriptive answers) are found to be work effectively in this system. The evaluation without disambiguation is done using the same set of questions in an online question answering system named 'answers.wikia.com'. It is found that disambiguation will increases the speed of retrieving the answers while comparing the retrieving speed of both online website and the proposed work eventhough there is only a slight increase in the precision, recall and f-measure. The reason for the increased speed is that, it will make a search after disambiguating the entities. In factoid queries, the questions starting with when and why are found to be not answered in this system. Since this system is based on the disambiguation of named entities and the question starting with why mainly represents an event or a consequence, this system will not retrieve answers. Also the questions starting with when is time related, but the time is a relative entity and not independent. So it comes with any other named entities person, location etc. In the case of factoid queries, short precise answers are retrieved. In the case of definition queries, long descriptive answers are retrieved.

V. Conclusion and Future Scope

The proposed work have introduced an integrated system that perform Speech Recognition, Named Entity Disambiguation, Anaphora Resolution and Question Answering system. Human speech is taken as input and converted into the text. It is then undergone Entity Disambiguation. Then Anaphora Resolution of named entities are performed and finally developed a question answering system that depends on the disambiguation of entities. The entity disambiguation is found to be more efficient if anaphora is performed along with it. Entities that are not correctly recognized in the disambiguation phase are found to be recognized, after the anaphora resolution. The question answering system is also found to be more efficient, because the disambiguation will improves the speed of answer retrieval than those systems without disambiguation. The anaphora resolution is implemented more better when it is integrated with the entity disambiguation ie. Pronominal anaphora, definite noun phrase anaphora, one-anaphora. The disambiguation with anaphora is more efficient than disambiguation without anaphora in terms of precision, recall and f-measure. Also the question answering system with disambiguation is more efficient than question answering system without disambiguation.

The future work will expect to accomplish all these tasks:

- Mapping the correct anaphors to the multiple antecedents.
- Disambiguation and recognition of the local names (ie. Not the name of a famous personality).

References

[1] Mena B. Habib, Maurice van Keulen , "NEED4Tweet: A Twitterbot for Tweets Named Entity Extraction Disambiguation", Proceedings of ACL-IJCNLP 2015 System Demonstrations, pp. 31- 36, 2015.

[2] Danica Damljanovic and Kalina Bontcheva "Named Entity Disambiguation using Linked Data", In proceedings of the 9th Extended Semantic Web Conference ESWC2012, pp. 334-336, 2012.

[3] Lev Ratinov, Dan Roth, Doug Downey, Mike Anderson "Local and Global Algorithms for Disambiguation to Wikipedia", In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pp. 1375-1384, 2011.

[4] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, Tim Finin, "Entity Disambiguation for Knowledge Base Population", COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, pp. 277-285, 2010.

[5] Vipin Kumar, Pandey Shreya, Solanki Kamini Sharma, "A Technique for Anaphora Resolution of Text", In proceedings of the International Journal of Applied Information Systems IJAIS, volume 5, pp. 0868-2249, 2013.

[6] Eugene Charniak and Micha Elsner, "EM Works for Pronoun Anaphora Resolution", In Proceedings of the 12th Conference of the European Chapter of the ACL EACL 2009, pp. 146-158, 2009.

[7] Kalyani P. Kamune, Avinash Agrawal, Hybrid Approach to Pronominal Anaphora Resolution in English Newspaper Text, International Journal of Intelligent Systems and Applications(IJISA), vol 7, pp. 56-64, 2015.

[8] Smita Singh, Priya Lakhmani, Dr. Pratistha Mathur, Dr.Sudha Morwal, Analysis of Anaphora Resolution System for English Language, In proceedings of International Journal on Information Theory (IJIT), vol 3, pp. 51-57, 2014.

[9] Niyu Ge, John Hale, Eugene Charniak, A Statistical Approach to Anaphora Resolution, In proceedings of the workshop on very large corpora, pp. 161-170, 1998.

[10] Gayatri Chavan, Sonal Gore, Design of the Effective Question Answering System by Performing Question Analysis using the Classifier,International Journal of Computer Applications Foundation of Computer Science (FCS), NY, USA, vol 139, pp. 1-3, 2016.

[11] Marcus Klang, Pierre Nugues, Named Entity Disambiguation in a Question Answering System, The Fifth Swedish Language Technology Conference (SLTC 2014), 2014.

[12] Rohini Srihari, Wei Li, A Question Answering System Supported by Information Extraction, In Proceedings of the sixth conference on Applied natural language processing, pp. 166-172, 2000.

[13] Taniya Mishra, Srinivas Bangalore, Qme! : A Speech based Question-Answering system on Mobile Devices, In Proceeding HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 55-63, 2010.