Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Importance of retrieving noun phrases and named entities from digital library content

Ratna SANYAL, Kushal KESHRI, Vidya NAND

(*Indian Institute of Information Technology, Allahabad 211012, India*)

E-mail: {rsanyal, iit2006031, iit2006032}@iiita.ac.in

Received Sept. 1, 2010; Revision accepted Sept. 16, 2010; Crosschecked Sept. 1, 2010

Abstract: We present a novel approach for extracting noun phrases in general and named entities in particular from a digital repository of text documents. The problem of coreference resolution has been divided into two subproblems: pronoun resolution and non-pronominal resolution. A rule based-technique was used for pronoun resolution while a learning approach for non-pronominal resolution. For named entity resolution, disambiguation arises mainly due to polysemy and synonymy. The proposed approach fixes both problems with the help of WordNet and the Word Sense Disambiguation tool. The proposed approach, to our knowledge, outperforms several baseline techniques with a higher balanced *F*-measure, which is harmonic mean of recall and precision. The improvements in the system performance are due to the filtering of antecedents for the anaphor based on several linguistic disagreements, use of a hybrid approach, and increment in the feature vector to include more linguistic details in the learning technique.

Key words: Coreference resolution, Hybrid approach, Filtering, Rule based and J48 algorithm
doi:10.1631/jzus.C1001003 **Document code:** A **CLC number:** TP391

1 Introduction

In this world of an ever-increasing repository of digital libraries, it has become very important to be able to extract structured and useful information. The creation and management of the digital library system requires the development of software that allows the acquisition of data for the digital library, and allows their storage, indexing, and subsequent retrieval based on the requests of the users. Coreference resolution is an important research area in this regard. Coreference, as the name indicates, is said to occur when multiple phrases refer to the same real world entity.

For example, in the sentence “Yesterday I played football with John Miller. He really impressed everyone with his performance”, ‘John Miller’, ‘He’, and ‘his’ are most likely coreferent to the same person John. Thus, they belong to the same coreferent class. Although we as readers have little problem relating the noun phrases to the appropriate referent class, the same is not true for natural language processing systems. Coreference resolution presents a formidable challenge.

Coreference resolution can be divided into two classes based on appearance in an expression. Anaphora is used to refer to an expression that has occurred prior. In the earlier example, ‘he’ and ‘his’ are an instance of anaphora, and ‘John Miller’ is called its possible antecedent. Cataphora is an antonym of anaphora and is used to find a forward reference in a discourse. For example, “My friend, Robert, is a great scholar”. Here ‘My friend’ is an instance of cataphora and refers to ‘Robert’ appearing after it. Our proposed approach is applicable for anaphora resolution. Researchers previously marked their breakthroughs with a rule set to resolve coreference resolution, but a number of learning approaches have come into existence. Our approach is a hybrid one, using both learning and non-learning techniques, and

© Zhejiang University and Springer-Verlag Berlin Heidelberg 2010

reworking baseline methods (Lappin and Leass, 1994; Soon *et al*., 2001; Ng and Cardie, 2002).

1.1 Non-learning approaches

An initial breakthrough in the field of coreference resolution using non-learning techniques was achieved by Hobbs (1978). The author proposed two approaches for pronoun resolution, the first being the naive algorithm, which resolves pronouns like he, she, they, etc. In this work, a syntactic parse tree for each sentence was built. With the help of the breadth-first approach, the possible antecedents were analyzed based on number and gender agreement. The second approach was directed at pronoun resolution in a system analyzing semantics of English text. Lappin and Leass (1994) took the work of Hobbs (1978) to the next level by associating the salience factor with each possible antecedent for the resolution of personal pronouns. The entire set of possible antecedents was collected, and assigned a weight according to the sub-process. After removing antecedents having different linguistic features compared to the pronoun, the best antecedent was selected based on the highest salience weight. A modified version of the Lappin & Leass algorithm was then proposed in Kennedy and Boguraev (1996). The Centering Theory (Grosz *et al*., 1995) for discourse made an interesting observation that certain anaphors are more likely to be resolvable by having knowledge about the discourse structure of text. Later, many more ideas were suggested by researchers (Mitkov, 2002; Poesio *et al*., 2002).

1.2 Learning approaches

One of the pioneering works (Dagan and Itai, 1990) was aimed towards removing the need to manually acquire semantic information. It is possible to collect details on co-occurrence patterns for large corpora automatically by using this approach. Another research work (Fisher and Ellen, 1992) came up with an idea of resolving relative pronouns based on a statistical approach using a corpus. The technique for resolving relative pronouns (Cardie, 1992) was based on the clustering technique. In this work, the system was trained with training data. The system behaved on the basis of the most similar data found in the training samples for any unseen data in the test cases.

A machine learning tool was developed by McCarthy and Lehnert (1995) and named RESOLVE for coreference resolution using a decision tree. RESOLVE produces better results than hand crafted rules on a collection of English newspapers. The decision tree for machine learning was also used in Aone and Bennett (1995). In this case, the authors used 66 features in contrast to only 8 semantic features in RESOLVE. This was targeted for Japanese text.

The first learning based system to have recall and precision comparable to non-learning approaches was described in Soon *et al*. (2001). Recall is defined as the measure of completeness and precision the measure of exactness. In our work, we will use the same accuracy measures as used by Soon *et al*. (2001) and other earlier researchers. This approach was also used for more feature vectors and for modifying the training approach as per Ng and Cardie (2002).

The best scoring system of the MUC-6 data corpus (Kameyama, 1997) is available, following the approach of pre-extraction coreference resolution. It was loosely based on the Lappin & Leass algorithm, in which the antecedent was collected, filtered, and then ordered by the salience value.

Our approach tries to rework baseline approaches (Lappin and Leass, 1994; Soon *et al*., 2001; Ng and Cardie, 2002). Our proposed approach is a combination of learning and non-learning techniques on the above mentioned methods, by extracting its best features and reworking them. We have also paid special attention to the named entities.

2 The proposed coreference system

Our proposed technique is a hybrid approach for resolving coreference resolution by dividing it into two subproblems: pronoun resolution and non-pronominal resolution. The first subproblem uses the algorithm proposed by Lappin and Leass (1994) for pronoun resolution. The second subproblem first uses the filtering of antecedents for anaphora. All the candidate antecedents of an anaphor are collected and then filtered based on number, gender, etc. disagreement with the anaphor. This reduces the probability of a false referent. Fig. 1 depicts the basic outline of our approach.

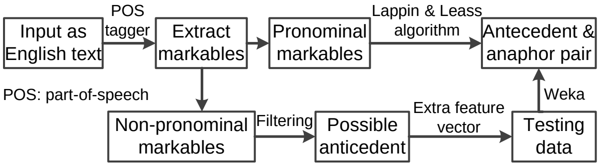

Fig. 1 Basic outline of our approach

The idea behind the proposed approach is inspired from studying baseline techniques (Soon *et al*., 2001; Ng and Cardie, 2002), in order to modify it for better accuracy. The drawbacks associated with these baseline approaches are that it uses a machine based approach for both pronoun and non-pronominal resolution. If we go into the details of the Discourse Theory (Mitkov, 2002), it is revealed that the local focus plays a very important role for the resolution of pronouns. Thus, the approaches followed by these baseline techniques (Soon *et al*., 2001; Ng and Cardie, 2002) are more prone to error in pronoun resolution. Hence, for pronoun resolution we are using a non-learning technique. Second, for the improvement of non-pronominal resolution, a few linguistic attributes have been added to improve the decision concerning whether an anaphor refers to an antecedent or not. Within non-pronominal resolution special attention has been given to named entity resolution. The approaches for both categories are described below.

2.1 Resolving pronouns

As in accordance with the factors associated with pronoun resolution, a local focus of the discourse should be taken into consideration. It can be verified from a manually annotated corpus that, in most cases, the antecedent of the anaphor lies within the first few lines before the anaphor (usually before 1st, 2nd, and 3rd lines). Thus, a non-learning technique has been used for pronoun resolution.

We follow the approach proposed in Lappin and Leass (1994). It can handle both inter- and intra-sentential pronoun resolution. Third-person pronoun, reflexive pronoun, and reciprocal pronoun are resolved with a very good accuracy with this approach. Pleonastic pronoun is also handled efficiently using this method. The input for the tool is English text documents. At first, all the markables are extracted. These markables can be pronoun-phrase, and non-pronominal phrase including a named entity. Extraction of markables is done by the preprocessing of English text with a part-of-speech (POS) tagger. After the preprocessing for a markable M_x (called ‘anaphor’), we look for the markable M_y appearing before the anaphor in the text (called ‘antecedent’), for choosing the best referring antecedent. For intra-sentential pronoun resolution a syntactic filter is used.

The anaphor binding algorithm is used for determining the possible antecedents for the resolution of reflexive and reciprocal pronouns. Now the possible antecedents are assigned a salience weight based on several salience factors such as sentence recency, subject emphasis, and head noun emphasis. Among all the possible antecedents for an anaphor, the one having the maximum salience weight is chosen as the coreferent antecedent. The entire approach is based on the fact that a pronoun always corefers to the antecedent, except for pleonastic pronouns, which have been dealt with differently with the help of syntactic and partially lexical techniques. Specification of modal adjectives such as necessary, possible, etc. and cognitive verbs such as recommend, believe, etc. help the tool to recognize pleonastic pronouns.

2.2 Resolving other noun phrases

The learning technique has been followed for resolving other noun phrases including named entities. For this subproblem we use a similar approach to what we have used for pronoun resolution, by extracting all the markables obtained from preprocessing of text. For an anaphor markable M_x, we collect all the antecedents markables M_y appearing before it in the text document, and filter it on the basis of number, person, and gender disagreement to reduce the false referent. Table 1 depicts the filtering process with decision true, if they are not coreferent, otherwise false, which adds M_y to the possible sets of antecedent which can corefer to anaphor.

Table 1 Filtering process

Attribute	Decision (true or false)
Number disagreement	True if M_x and M_y do not agree in number [e.g., Dog and Dogs do not agree in number]
Gender disagreement	True if M_x and M_y do not agree in gender [e.g., Mr. Robert & Mrs. Robert do not agree in gender]
Semantic class disagreement	True if M_x and M_y do not share the same semantic class [e.g., New York (place) and New York Times (organization) do not agree in semantic manner]

After taking out the possible set of antecedents for an anaphor through filtering, the machine learning tool Weka (Reutemann *et al*., 2004) was used to resolve other noun phrases. The algorithm used for this purpose is the greedy J48 algorithm, which is based on the C4.5 decision tree learner, which learns from training data. Training data (http://afner.sourceforge.net/) is prepared from several online annotated corpora using the method specified below.

2.2.1 Preparing training data for learning

From the annotated corpora, we extract attributes between possible antecedent and anaphor with the help of tools and techniques specified in the third section. Table 2 depicts all the possible attributes which are required for preparing the training set.

For generating training examples, the modified Soon's approach (Soon *et al*., 2001) has been used. Suppose c1-c2-c3-c4 is a coreference chain in annotated corpus. Then for creating positive samples we look for non-pronominal antecedents as we have solved the pronoun resolution separately. For example, if c1, c3, c4 are other noun phrases and c2 is pronoun, then the positive sample consists of c1-c3, c3-c4, and c1-c4. For generating negative training examples in between c3 and c4, there are two other non-pronominal phrases such as a, b, and then the negative training samples with c4 anaphor will be a-c4, b-c4. In this way, training data is prepared and the machine tool is trained. After training the tool, the learning model is saved for further use. For testing data we use the saved model for making the decision of whether two markables corefer or not. The decision is made by the decision tree in a top-down fashion.

Testing samples are similarly created as training samples by extracting different attributes. Testing samples are now filtered so that we send a smaller number of data samples to the machine for making the decision. An antecedent-anaphor pair is not coreferent if they can be filtered based on number disagreement, gender disagreement, and semantic class disagreement as specified in Table 1.

Semantic class agreement is an important attribute for name entity resolution. In the proposed approach, we have taken entities like place, location, organization, date, time, and monetary values into consideration. For resolution of named entity, polysemy and synonymy also play an important role. A word can have different meanings in different contexts. With the help of the word sense disambiguation process, the system determines whether the markables are used as noun or verb, etc., and then with the help of WordNet (http://wordnet.princeton.edu), the meaning in the given context is evaluated.

Table 2 All possible attributes for the training set

Attribute	Value
String match	If the markables M_x and M_y match, the value is True; else False
Same head	If M_x and M_y have a similar head, the value is True; else False. The head of the noun phrase is the word that determines the semantics of the phrase and its syntactic type
Definite	True if M_x noun phrase is preceded with the article 'the'; else False
Demonstrative	True if M_x noun phrase is preceded with 'this', 'that', 'these', and 'those'; else False
Number agreement	Its values are True, False, or Unknown. If M_x and M_y agree in number, then the value is True; if not, the value is False. In few cases we cannot decide whether they agree or not, the value is unknown
Both proper noun	True if M_x and M_y both fall into the proper noun category; else False
Part of	True if either one of them (M_x or M_y) is part of the other; else False
Semantic class agreement	Its values are True, False, or Unknown. True if M_x and M_y share the same semantic class; False, if they share different semantic classes. In ambiguous cases the value is Unknown
Abbreviation (Aliasing)	True if M_x is alias of M_y; otherwise False. [I.B.M. is an acronym used for International Business Machines]
Distance features	Its values are 0, 1, 2, ... Values are based on the appearance of M_x and M_y in the text documents. If they appear in the same line, the value is 0; if they are one sentence apart, the value is 1; and so on
Gender	Its values are True, False, or Unknown. True if M_x and M_y agree in gender; False if they do not. In cases we cannot decide the gender the value is Unknown
Appositive	True if M_x is in apposition to M_y; otherwise False [e.g., My friend John is presently working at a firm in New York. Here 'John' is in apposition to 'My friend']

After filtering the markables, the remaining testing data is sent to the machine tool for making a decision. The output of the machine tool is either true or false with a certain probability. The output of the tool is calculated by using the decision tree. After pre-filtering, we once again include the features such as gender, number agreement, and semantic class agreement to the feature vector.

This is done just to increase the information gain of the decision tree built using the training data. Out of the possible 11 attributes, the root of the decision tree is decided by the factor of information gain. The attribute with the highest information gain is selected as a root and then the tree is parsed in a top-down fashion starting from the root, in order to calculate the decision of whether a pair of M_x and M_y corefer or not.

3 Experimental setup

3.1 Tools and techniques

3.1.1 Parser

For the named entity resolution, we use the POS tagged text. For pronoun resolution we have to look into the grammatical structure of sentences such as subject, object, etc. The system uses a statistical parser developed by Stanford University, known as the Stanford Parser (http://nlp.stanford.edu/) to accomplish this process.

According to the requirements, the parser can generate output in various formats for the plain English text. Among the three different parsers available, the system uses the lexicalized probabilistic context free grammar (PCFG) parser for processing the text. The parser handles the new sentence based on the knowledge gained from the hand parsed sentences.

3.1.2 WordNet 2.1

WordNet is highly useful for determining the semantic properties of the noun phrases. We can track down the antecedent M_y that has the same semantic sense as the anaphor M_x. All the synsets of the word are mentioned in order of their estimated frequency and therefore we consider only the sense with the highest frequency. Hypernyms give us the general class of the noun phrase to which it belongs, whereas hyponyms give us the most specific class of the noun phrase.

3.1.3 Weka

Waikato Environment for Knowledge Analysis (Weka) (Kameyama, 1997; Soon *et al*., 2001; Ng and Cardie, 2002) is a machine learning tool which can learn from annotated training data. Weka contains a large collection of algorithms for data mining. For our purpose we used the J48 learning algorithm.

3.1.4 Named entity recognition

An efficient recognition of named entities plays a vital role in improving the accuracy of the system. Named entity recognition helps to classify noun phrases into certain predefined categories such as name of person, location, organization, identification of quantitative value, date, etc. We can classify the books, documents, etc. by identifying the same noun phrases or named entities. The Stanford named entity recognizer (Soon *et al*., 2001) is the most efficient tool for recognizing named entities, but it handles only the name of person, location, and organization. For the rest of the types, our system uses another named entity recognition tool (http://www.cs.waikato.ac.nz/).

3.2 Gender database

Determining correctly the gender of the antecedent and anaphor is very important for coreference resolution. This is evaluated with the help of a gender database. However, in a few cases ambiguity arises due to several noun phrases having different genders in different contexts. The database used by the system was built using a large repository of online text (Bergsma and Lin, 2006).

The above system has the following format:

Noun phrase *tab* Masculine_Count *space* Feminine_Count *space* Neutral_Count *space* Plural_ Count.

As we can observe in the format of the database, a noun phrase is followed by the count of different gender forms. Thus, the one having the maximum count among masculine, feminine, or neutral will be the gender of the noun phrase in question. If we cannot determine the gender of the noun phrase with the help of the database, then the system treats it as unknown.

4 Results and analysis

All the baseline approaches (Lappin and Leass, 1994; Soon *et al*., 2001; Ng and Cardie, 2002) have been either learning or rule based approaches. A hybrid approach has not been used till now. Soon *et al*.

(2001) made analysis on MUC-6 and MUC-7 using the learning approach. The recall is 41.4% and precision is 59.1%. The learning approach in Ng and Cardie (2002) improves upon the approach of Soon *et al*. (2001) by a significant margin.

For non-pronominal resolution we followed the approach discussed above. The improvement has been mainly due to the filtering of antecedent, minimizing the chances of a false referent before passing it to the machine tool for decision, adding extra linguistic features, and modifying the training approach. The recall of the system is 59%, while the precision is 67%, resulting in an *F*-measure of 62.74%. In this way, by classifying the anaphora resolution task into two subproblems we can improve the overall accuracy.

5 Conclusions

As mentioned, we can observe the drawback associated with the baseline approach and the proposed approach worked upon it to improve recall and precision. Thus, we have used the pre-filtering and learning approach for named entities, and the Lappin & Leass approach for pronoun resolution. We have improved the system accuracy by first filtering the dataset and then passing the refined dataset for further processing. Extra linguistic features also contribute to this improvement.

A digital repository becomes a rich resource for researchers to solve the issues related to a digital library. To obtain the information on similar topics, the same author, related documents, etc. from the digital library, the extraction of noun phrases and named entities can be used as one of the parameters. There are born digital data and digitized documents in all over the world which are collected in the different digital repositories. Thus, we have many documents in different languages on the same person, place, year, etc. The research on the retrieval of noun phrases and named entities in different languages will resolve this multilingual problem. This work is in progress.

Acknowledgements

We gratefully thank the Indian Institute of Information Technology Allahabad and Ministry of Communication and Information Technology, New Delhi, Government of India for their support during our research.

References

Aone, C., Bennett, S.W., 1995. Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies. Proc. 33rd Annual Meeting on Association for Computational Linguistics, p.122-129.

Bergsma, S., Lin, D., 2006. Bootstrapping Path-Based Pronoun Resolution. Proc. Conf. on Computational Linguistics, p.33-40.

Cardie, C., 1992. Learning to Disambiguate Relative Pronouns. Proc. 10th National Conf. on Artificial Intelligence, p.38-43.

Dagan, I., Itai, A., 1990. Automatic Processing of Large Corpora for the Resolution of Anaphora References. Proc. 13th Int. Conf. on Computational Linguistics, **3**:1-3.

Fisher, D., Ellen, R., 1992. Applying Statistical Methods to Small Corpora: Benefitting from a Limited Domain. Probabilistic Approaches to Natural Language, Technical Report FS-92-05. American Association for Artificial Intelligence, AAAI Press.

Grosz, B.J., Joshi, A.K., Weinstein, S., 1995. Centering: a framework for modeling the local coherence of discourse. *Comput. Ling.*, **21**:203-226.

Hobbs, J.R., 1978. Resolving pronoun references. *Lingua*, **44**(4):311-338. [doi:10.1016/0024-3841(78)90006-2]

Kameyama, M., 1997. Recognizing Referential Links: an Information Extraction Perspective. Technical Report, AI Center, SRI International.

Kennedy, C., Boguraev, B., 1996. Pronominal Anaphora Resolution without a Parser. Proc. 16th Int. Conf. on Computational Linguistics, **1**:113-118.

Lappin, S., Leass, H.J., 1994. An algorithm for pronominal anaphora resolution. *Comput. Ling.*, **20**(4):535-561.

McCarthy, J.F., Lehnert, W.G., 1995. Using Decision Trees for Coreference Resolution. Proc. 14th Int. Joint Conf. on Artificial Intelligence, p.1050-1055.

Mitkov, R., 2002. Anaphora resolution. *Comput. Ling.*, **29**(4).

Ng, V., Cardie, C., 2002. Improving Machine Learning Approaches to Coreference Resolution. Proc. 40th Annual Meeting of the Association for Computational Linguistics, p.104-111.

Poesio, M., Ishikawa, T., Im Walde, S.S., Vieira, R., 2002. Acquiring Lexical Knowledge for Anaphora Resolution. Proc. 3rd Conf. on Language Resources and Evaluation, p.1220-1224.

Reutemann, P., Pfahringer, B., Frank, E., 2004. A Toolbox for Learning from Relational Data with Propositional and Multi-Instance Learners. 17th Australian Joint Conf. on Artificial Intelligence, p.1017-1023.

Soon, W.M., Ng, H.T., Lim, D., 2001. A machine learning approach to coreference resolution of noun phrase. *Comput. Ling.*, **27**(4):521-544. [doi:10.1162/089120101753342653]