Two-Stage Bootstrapping for Anaphora Resolution

Balaji J, T V Geetha, Ranjani Parthasarathi, Madhan Karky

Dept. of CSE & IST,

Anna University, Chennai - 600 025

Abstract:

In this paper, we propose a two-stage bootstrapping approach to resolve various anaphora representing persons, places, plurals and events in Tamil text. The existing approaches dealt with only single pronoun type and not all anaphora using common approach. Moreover, most of the approaches concentrate on syntax- based algorithms and semantics to some extent. Instead in our approach, we tackle various types of pronouns using a semi-supervised bootstrapping approach with uniform pattern representation and by exploring the semantic features in resolving anaphors. In order to aid the semantics, we use Universal Networking Language (UNL), a deep semantic representation for resolving various types of pronouns. The two stages of our bootstrapping approach consists of identification of anaphora and its set of referring expressions in stage 1 and identification of correct antecedent of a pronoun in stage 2. In our approach two patterns are defined – one for anaphora and other for set of referring expressions. In addition, we introduce triggering tuples, which can be word based semantics or context based semantics, represented in the pattern of both anaphora and referring expressions so as to resolve the ambiguities during the identification of correct antecedent. The performance of our bootstrapping approach gives better results and proved.

Keywords: Anaphora resolution, Universal Networking Language, Bootstrapping, Triggering Tuples

Proceedings of COLING 2012: Posters, pages 507–516,
COLING 2012, Mumbai, December 2012.

1. Introduction

Anaphora Resolution commonly called Pronoun resolution is a problem of finding references in the previous utterances of a pronoun. The references can be noun, noun phrase, verb phrase and/or clause. The main aim of anaphora resolution is to find the correct antecedent of a pronoun from the set of referring expressions. The antecedent of a pronoun is identified by the set of features such as number gender agreement features, grammatical relations for person pronouns and verb predicates for plural and event pronouns.

Popular syntax-based approaches include Centering theory (Brennan et al, 1987) and Hobb's algorithm (Hobbs, 1978) in which the both the algorithms are used to resolve person pronouns using agreement features and grammatical ordering of relations. The verb predicate feature has been used to identify event pronouns using a composite kernel method (Bin et al, 2010). In addition, rule-based has been attempted to resolve personal pronouns. The rules do not fit to resolve entire pronoun resolution task. In contrast, machine learning approaches have also been attempted to resolve person pronouns automatically. In particular, most of these techniques dealt with resolving only person pronouns and in some cases plurals and event pronouns (Chen et al, 2009).

In this paper, we propose a two-stage bootstrapping approach to resolve anaphora representing person, place, plural and events automatically from Tamil text. We define a uniform pattern to detect all types of pronouns while for referring expressions we define two types of patterns, one to tackle person and place pronouns and other to tackle plural and event pronouns. To our knowledge, this is the first attempt to tackle all types of pronouns using a single bootstrapping framework. We introduce the concept of triggering tuples in both the patterns of anaphora and its corresponding referring expressions to identify the correct antecedent. In order to aid the semantic compatibility information in anaphora resolution, we use Universal Networking Language (UNL) (UNDL, 2012), a deep semantic representation, to represent the referring expressions in the form of directed acyclic graphs. In this paper, the word level semantics and context level semantics information are utilized to resolve all the pronoun types. In addition, in order to generate new patterns, the semantic similarity of the antecedents is measured using the taxonomy of semantic constraints called UNL Ontology (UNDL 2012). While in our previous rule-based approach, we use three UNL relation based rules to tackle plural and event pronouns, however in this work, we have generalized the semantic relations to tackle more number of instances.

The paper is organized as follows. Section 2 discusses the related works on pronoun resolution. Section 3 describes the semi-supervised learning – bootstrapping of pronoun resolution which includes features, pattern representation and different stages of bootstrapping in finding the antecedent of a pronoun and generation of new patterns. In section 4, we discuss the performance of our approach and compared with another bootstrapping system. Finally, we conclude our approach with future enhancements.

2. Related Work

In this section, we discuss various techniques attempted for resolving different types of anaphora. A modified Centering theory and a rule-based approach has been proposed for resolving pronouns such as person, place, plural and events in which the rules are based on word-level semantics such as semantic constraints and sentence-level semantics such as UNL relations

(Balaji[2] et al, 2011). A robust rule based system has been applied to resolve personal pronouns, subject and dative pronouns in French in which the rules are based on agreement features and syntactic structure (Trouilleux, 2002). A syntactic rule based Hobb's algorithm has been used to resolve possessive and reflexive pronouns of Hindi language (Kamlesh et al, 2008). In contrast to the rule based approaches, machine learning approaches such as conditional random field (CRF) statistical model for chinese(Li & Shi, 2008), Tamil (Murthi et al, 2007) have been attempted. A twin candidate based learning model has been proposed to resolve event pronouns in English in which the verb predicates are considered as antecedents (Bin et al, 2010).

In the existing approach (Balaji[2] et al, 2011), various pronoun types have been resolved using a set of rules. However, rules have limited knowledge and do not provide accurate results for large set of sentences. In contrast, machine learning techniques focused on using syntactic features. to resolve person pronouns and co-reference chains in which the patterns defined are based on syntactic paths and word associations. Moreover, gender/number information and semantic compatibility have been determined using a probabilistic behavior (Bergsma et al, 2006). Another bootstrapping procedure for co-reference resolution uses word association information and are labeled using a self-trained approach (Kobdani et al, 2011).

However, the use of syntactic features and paths are difficult in relatively free-word order languages like Tamil and the approaches described above consider only limited types of anaphora. In order to overcome these difficulties, we propose a semi-supervised, two-stage bootstrapping procedure to resolve person, place, plural and event pronouns. The input to our bootstrapping framework is a set of UNL semantic graphs. The word level and context level information obtained from UNL graphs is utilized to define the example patterns. We introduce various scoring schemes during the filtering of non-referring expressions, choosing the correct antecedent and, the confidence of tuples and dependency relations in resolving pronouns. In addition to the scoring, we measure the semantic similarity between the semantic tuples of the referring expressions and propose a generalized procedure to learn new set coordinating and subordinating UNL relations to find the correct antecedent of a pronoun along with the existing relations proposed in the existing work (Balaji[2] et al, 2011).

3. Bootstrapping for Anaphora Resolution

Bootstrapping is a task of iteratively learning new patterns from unlabeled data, starting with a small labeled data from which the seed patterns are obtained. In this paper, we describe a two stage bootstrapping approach to resolve various types of anaphora. Our bootstrapping approach consists of two stages.

Stage 1: Extraction of anaphora and an associated set of referring expressions

Stage 2: Identification of the correct antecedent of the anaphora from the referring expressions obtained from Stage 1.

3.1 Features used for pattern representation in Anaphora Resolution

The pattern in the anaphora resolution is represented using the word-based features and context-based features. We use two classes of features, one for detecting anaphora and the other to extract the corresponding referring expressions. Here, each class consists of both word-based and context-based features. In addition, we introduce the concept of triggering tuple which forms part of the pattern representation but however does not take part in the actual matching process.

The triggering tuple helps the bootstrapping process to select the correct antecedent of a pronoun from the set of referring expressions obtained for that pronoun. The features are listed in Table-1.

Features for representing Referring Expressions
POS of the word (W_i) (Nouns or Entities) - POS(W_i)
Semantic Constraint associated with the word (W_i) - SC(W_i)
Attributes associated with the word (W_i) - ATTR(W_i)
Semantic Relation connected with the word (W_i) - SR(W_i)
Features for representing Anaphora
Pronoun (P_j)
POS of Pronoun (P_j) - POS(P_j)
Type of Pronoun (P_j) - PT(P_j)
Attributes associated with Pronoun (P_j) - ATTR(P_j)
Semantic Relation connected with Pronoun (P_j) - SR(P_j)
Triggering tuples
Verb (V) - Verb(V)
Attributes associated with the verb (V) - ATTR(V)
Semantic Constraints associated with the verb (V) - SC(V)
Attributes associated with the word (W_i) and Pronoun (P_j)
Semantic Relation connected with the word (W_i) and Pronoun (P_j)

TABLE 1 Features for resolving various types of anaphora

From Table-1, we introduce a new concept called triggering tuples which are used to signal the correct antecedent of a pronoun from the detected set of referring expressions. The triggering tuples are represented in the patterns of both referring expressions and pronouns. The triggering tuples can be a set of word based features and context based features

3.2 Pattern representation

In our approach, patterns are defined with use the graph based features (including both word-based and context-based features. The pattern representation for the set of referring expressions corresponding to anaphora representing persons and places is different from the referring expressions corresponding to anaphora representing plurals and events and thus can be shown in table 2. However, it is to be noted that the pattern representation for anaphora is common to all anaphora types. The pattern representations are described in detail in the following sections.

3.2.1 Pattern representation for Anaphora

The pattern representation is generic to all types of anaphora. Based on the features mentioned in Table-1, the pattern of anaphora is defined as

<Pronoun(W_i)+POS(W_i)+SC(W_i)+ATTR(W_i)+SR(W_i)-[Verb(V)+ATTR(V)+SC(V)]>

where j = 1,2 ….N

3.2.2 Pattern representation for a set of Referring Expressions

As described earlier, the pattern representation for the set of referring expressions corresponding to anaphora representing persons and places is different from the referring expressions of plural and event anaphora which are shown in table 2.

Referring Expressions for	Pattern Representation
Anaphora representing Person and Place	**$<POS(W_i)+SC(W_i)+ATTR(W_i)+SR(W_i)-[Verb(V)+ATTR(V)+SC(V)]>$** where i= 1, 2, 3 …N
Anaphora representing Plural and Events	**$<\{POS(W_i)+SC(W_i)+ATTR(W_i)+SR(W_i)-POS(W_k)+SC(W_k)+ATTR(W_k)+SR(W_k)\}_L-[Verb(V)+ATTR(V)+SC(V)]>$** where i, k, L = 1, 2, 3 …N & i~=k

TABLE 2 Pattern representations for referring expressions of various Pronouns

3.3 Stage 1: Extraction of Anaphora and associated referring expressions

During this stage, anaphora and the associated possible set of referring expressions are identified and extracted. The referring expressions are extracted based on the triggering tuple represented in the pattern of anaphora. Moreover, in order to reduce the redundant expressions such as non-referring expressions, a scoring function is introduced. This scoring is used to filter out non-referring expressions from the set.

3.3.1 Filtering of non-referring expressions

One of the important tasks of anaphora resolution is to filter out non-referring expressions that do not take part in resolving the anaphora type. The referring expressions of a pronoun can be nouns and entities. The non-referring expressions are usually filtered out using grammatical relations (Brennan et al, 1987). Instead of using the grammatical ordering for filtering out non-referring expressions, we use a scoring function for filtering. The scoring of referring expressions is based on the number of entities and/or nouns identified as referring expressions for the corresponding pronoun among the total number of referring expressions along with the type of anaphora to be resolved.

$$\text{Filter } RE_{As} = \frac{N\ RE_I + A_s}{RE_I + A_n}$$

Where RE_{As} - set of referring expressions for a specific anaphora, A_n – all types of anaphora, RE_I - referring expressions where I = 1, 2, 3 …N, A_s - specific anaphora to be resolved

Filtering of non-referring expressions can be performed by identifying the occurrence of a referring expression for a specific anaphora among the total occurrence of that regular expression with other anaphora types.

3.4 Stage 2: Identification of correct antecedent of a pronoun

During this stage, the correct antecedent of a pronoun is identified from the set of referring expressions obtained from stage 1 through the selection of complete patterns and partial patterns. The defined patterns represent the anaphora and the set of referring expressions. In order to choose the correct antecedent of a pronoun, the triggering tuples defined in the pattern are exploited using word based features and context based features.

3.4.1 Triggering Tuples

One of the important aspects of our bootstrapping approach is the use of triggering tuples to identify the correct antecedent of a pronoun. The idea behind the use of triggering tuples is to filter out the non-referring expressions among the set of referring expressions of a pronoun. In some cases, most approaches as discussed earlier failed to identify the correct antecedent of a pronoun. This is because the most popular algorithms such as Centering theory (Brennan et al, 1987), Hobb's algorithm (Hobbs, 1978) etc used in the existing approaches for resolving pronouns are syntax-based algorithms and not focused on the semantics of the word and/or context of a pronoun. However, Bin et al (2010) modified the centering theory by incorporating the semantic roles to resolve the pronouns. This approach solves the problem to an extent but only for person pronouns and the problem remains unsolvable in ambiguous cases (i.e. choosing an antecedent is ambiguous). Instead in our approach, this problem can be resolved by examining word based semantics and context based semantics. The following section describes the selection of correct antecedent of a pronoun.

3.4.2 Selection of complete patterns in identifying the antecedents of a corresponding pronoun

Identifying the correct antecedent of a pronoun is achieved by the selection of complete patterns. The triggering tuples of the patterns of anaphora and its referring expressions play a vital role in identifying the correct antecedent of a pronoun. A probabilistic scoring of triggering tuples in the pattern of anaphora and its corresponding referring expressions set is determined. Based on the scoring, confidence of the instances are examined and identified as an antecedent of a corresponding anaphora. The scoring of tuples is given below.

$$\text{Select } AN_{As} = \frac{N\ T_{RE_I + A_s}}{T_{RE_I + A_n}}$$

where AN_{As} – Antecedent of a specific anaphora, A_s - specific anaphora to be resolved, A_n – all types of anaphora, T – Triggering tuples, RE_I - referring expressions where I = 1, 2, 3 …

3.4.3 Selection of partial patterns to generate new patterns

After the complete pattern matching is performed, instances that are not tackled under exact matching are then partial matched. Partial matching is carried out at different levels. The first level is to modify the word-based features such as semantic constraints of referring expressions to obtain the semantic similarity between the example patterns of the referring expressions and the input instances of the referring expressions. The semantic similarity of semantic constraints is achieved by using the semantic UNL Ontology (UNDL, 2012) in which the semantic constraints are arranged in a hierarchal relations such as "is-a" and "instance-of". The next level of partial matching is performed at the context-based features such as UNL relations. In addition, the confidence of the tuple values is also computed to identify the correct antecedent of a pronoun.

The detailed analysis of semantic similarity of constraints and, coordinating and subordinating relations are described below.

3.4.3a Semantic Similarity of Constraints

The semantic similarity between constraints is useful in identifying the anaphora representing places. As discussed above, adverbs such as "ingu" and "angu" along with its morphological variations can act as pronouns which represent places. Semantic constraints such as city, town, state, country etc all represent places. However it is difficult to list all these semantic variations in a pattern. Instead a semantic abstraction of constraints is needed to tackle these variations. This is achieved by using the semantic abstraction of the semantic constraints which is available through the semantic UNL ontology (UNDL, 2012). Semantic similarity is measured by the distance between the parent semantic constraints in UNL Ontology and is given below.

$$SIM\ C_i, C_j = DIST\ C_{parent}\ C_i, C_j$$

Where C_i – Semantic Constraint of referring expression obtained for an instance, C_j – Semantic constraint of referring expression in an example pattern, C_{parent} – Parent semantic constraint in UNL Ontology, DIST – Distance measure

Next, we will discuss the handling of coordinating and subordinating relations necessary to obtain the correct antecedent of a pronoun.

3.4.3b Coordinating and Subordinating UNL Relations

UNL relations obtained for referring expressions that exactly matches with the UNL relation obtained for anaphors are coordinating UNL relations and UNL relations obtained for anaphors that infer the UNL relations obtained for referring expressions are subordinating UNL relations (Balaji et al, 2011). In addition, we explored more number of coordinating and subordinating relations to resolve various pronouns. The specific rules on UNL relations are also generalized and thus the antecedent of a pronoun can be decided by

1. Participant relations connected with pronoun that exactly matches with Participant relations connected with the referring expressions in the previous utterances. Here, the triggering tuple is a verb and its associated UNL attribute as unaccusative (NACC). Example shown below comes under this category and is resolved using the triggering tuples mentioned above. For example,

Ram baluvai adiththaan. Avan azhuthaan. Ramhit Balu. He cried.

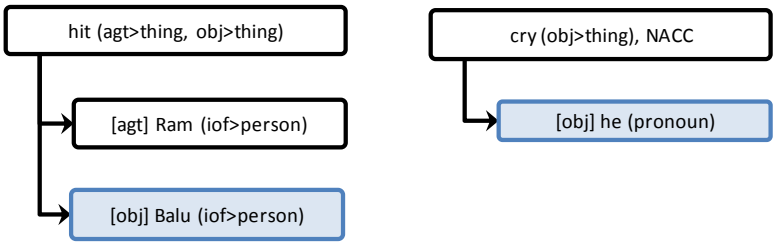

FIGURE 2 UNL Graphs for the sentences "Ram hit Balu. He cried".

The semantic constraints are shown in the braces and the relations connected with the corresponding concepts are shown in square brackets. The attribute "NACC" represents the verb is unaccusative.

The participant relation "obj" connected with pronoun that exactly matches with the participant relation "obj" connected with the concept "balu (iof>person)" in the previous sentence. Here, the triggering tuple is "NACC". And thus the antecedent of a pronoun "he" is identified as "balu". Similarly the other conditions are applied to obtain the correct antecedent of a pronoun.

2. Participant relations connected with pronoun that infers the Participant relations connected with the referring expressions in the previous utterances. Here, the triggering tuples are transitive verbs and its associated information.
3. Modifier relations connected with pronoun that infers the Participant relations connected with the referring expressions in the previous utterances. Here, the triggering tuples are transitive verbs.
4. Location relations connected with pronoun along with the "be" verb infers the Attribute relations connected with referring expressions in the previous utterances.

Using the conditions described above, the coordinating and subordinating relations are identified. In addition, new combination of relations is learned from the above conditions and thus the correct antecedent of a pronoun is obtained.

5. Evaluation

The performance of our bootstrapping approach is investigated using Tourism and News domain. We have considered 10000 sentences from each domain and tagged with the appropriate features such as POS, UNL attributes, UNL semantic constraints and UNL relation. We have taken 1000 tagged sentences for training data and extracted the most frequently occurred example patterns of each pronoun type. During this process we have identified 3025 person pronouns (singular), 2857 place pronouns, 156 plural pronouns and 323 event pronouns. Out of these obtained pronouns, we have achieved the overall result of 84% accuracy. The precision, recall and F-measure for resolving pronouns are shown in the table below. Table-3 shows the precision of various types of pronouns resolved.

Type of Anaphora	Precision	Recall	F-measure
Third Person Singular Pronouns	0.852	0.83	0.84
Third Person Plural Pronouns	0.79	0.74	0.76
Place pronouns	0.842	0.823	0.832
Event pronouns	0.837	0.656	0.735

TABLE 3 Performance of our Bootstrapping approach

We have also compared our bootstrapping approach with the previous rule based approach (Balaji et al, 2011). The comparison is shown in Fig 3. From the results, it is to be noted that our bootstrapping approach performs better than the previous rule based approach. Since the rules are limited in our previous approach, the F-measure is low and this difficulty could be resolved in our bootstrapping approach. We have also compared our bootstrapping approach with the path coreference of (Bergsama et al, 2006) which is a bootstrapping approach. The comparison of baseline system with our approach is shown in Fig 4. The parse tree for Tamil sentences is constructed using the existing Tamil parser (Saravanan et al, 2003). From the results, it can be seen that the performance of our approach gives better results than the existing system and the f-measure of our approach is 84% when compared to the baseline system as 48%.

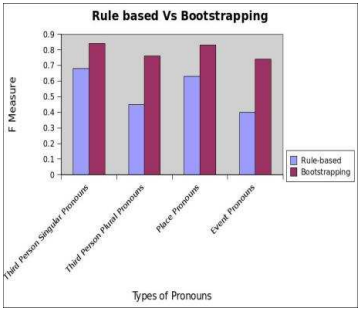

FIGURE 3 Comparison of Bootstrapping and Rule-based approach

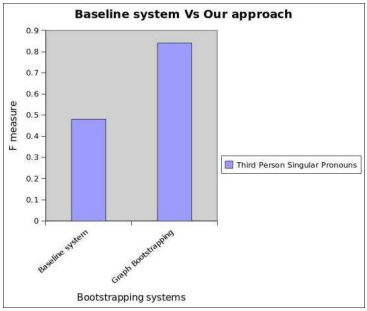

FIGURE 4 Comparison – Baseline system Vs. Graph Bootstrapping

Conclusion

In this paper, a semi-supervised, two-stage bootstrapping approach has been described to resolve all types of anaphora. In stage 1, the anaphora and its referring expressions are identified and in stage 2, the correct antecedent of a pronoun is selected among the set of referring expressions of a corresponding pronoun. This two-stage bootstrapping approach uses two patterns – for anaphora and referring expressions. Both the patterns consist of word based semantics and context based semantics. Moreover, a new concept called triggering tuples has been introduced in our bootstrapping approach so as to identify correct antecedent of a pronoun in case of ambiguities. The performance of our bootstrapping approach produces better results when compared to the baseline bootstrapping system. Further, we enhance this bootstrapping approach for identifying coreference entities by identifying more number of coordinating and subordinating relations.

References

Abney Steven, (2004), Understanding the yarowsky algorithm, Computational Intelligence, pages: 365-395

Balaji J[1], T V Geetha, Ranjani and Madhan Karky, (2011), Morpho-Semantic Features for Rule-based Tamil Enconversion, International Journal of Computer Applications 26(6):11-18, July 2011. Published by Foundation of Computer Science, New York, USA

Balaji J[2] and T. V. Geetha and Ranjani Parthasarathi and Madhan Karky, (2011), Anaphora Resolution in Tamil using Universal Networking Language, IICAI-11, pages: 1405-1415

Bergsma, Shane and Lin, Dekang, (2006), Bootstrapping path-based pronoun resolution, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, ACL-44, Sydney, Australia, pages: 33--40

Bin, Chen and Jian, Su and Lim, Tan Chew, (2010), A twin-candidate based approach for event pronoun resolution using composite kernel, Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10, Beijing, China, pages: 188—196

Brennan Susan E, Marilyn W. Friedman , Carl J. Pollard, (1987), A centering approach to pronouns, Proceedings of the 25th annual meeting on Association for Computational Linguistics, p.155-162, July 06-09, 1987, Stanford, California

Chen, Zheng and Ji, Heng and Haralick, Robert, (2009), A pairwise event coreference model, feature impact and evaluation for event coreference resolution, Proceedings of the Workshop on Events in Emerging Text Types, eETTs '09, pages: 17—22

Chen, Zheng and Ji, Heng, (2009), Graph-based event coreference resolution, Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, TextGraphs-4, pages: 54—57

Duan Manjuan and Jiang Ping, (2010), An Empirical study of Centering in Chinese Anaphoric Resolution, International Conference on Artificial Intelligence and Computational Intelligence, IEEE, pages: 373-377

Hobbs, J. R. (1978), Resolving Pronoun references, Lingua 44:311-338

Kamlesh Dutta, Nupur Prakash, and Saroj Kaushik, (2008), Resolving Pronominal Anaphora in Hindi using Hobbs' Algorithm, Web Journal of Formal Computation and Cognitive Linguistics, Vol. 1, No. 10

Kobdani, Hamidreza and Schütze, Hinrich and Schiehlen, Michael and Kamp, Hans, (2011), Bootstrapping coreference resolution using word associations, The Association for Computational Linguistics, ACL, pages: 783-792

Li Fei and Shi Shuicai, (2008), Chinese Pronominal Anaphora Resolution based on Conditional Random Fields, International Conference on Computer Science and Software Engineering, IEEE, pages: 732-734

Narayana Murthi, K.N, Sobha, L, Muthukumari, B. (2007), Pronominal Resolution in Tamil Using Machine Learning Approach, The First Workshop on Anaphora Resolution (WAR I), Ed Christer Johansson, Cambridge Scholars Publishing, 15 Angerton Gardens, Newcastle, NE5 2JA, UK pp.39-50

Saravanan K, Ranjani Parthasarathi and T V Geetha, (2003), Syntactic Parser for Tamil, In INFITT-2003

Shalom Lappin , Herbert J. Leass, (1994), An algorithm for pronominal anaphora resolution, Computational Linguistics, v.20 n.4, p.535-561, December 1994

UNDL, Universal Networking Digital Language, (2012), http: //www.undl.org/ Online; accessed 28 Jan (2012)