Hybrid Approach for Coreference Resolution

First Author: Sobha, Lalitha Devi., Pattabhi, RK Rao., Vijay Sundar Ram, R.
Second Author: Malarkodi, CS., Akilandeswari, A.
AU-KBC Research Centre,
MIT Campus of Anna University,
Chrompet, Chennai, India.
sobha@au-kbc.org

Abstract

This paper describes our participation in the CoNLL-2011 shared task for closed task. The approach used combines refined salience measure based pronominal resolution and CRFs for non-pronominal resolution. In this work we also use machine learning based approach for identifying non-anaphoric pronouns.

1 Introduction

In this paper we describe our system, used in the CoNLL-2011 shared task "Modeling Unrestricted Coreference in OntoNotes". The goal of this task is to identify coreference chains in a document. The coreference chains can include names, nominal mentions, pronouns, verbs that are coreferenced with a noun phrases.

The coreferents are classified into two types, pronominal and non-pronominal referents. We use two different approaches using machine learning and salience factor in the resolution of the above two types. Pronominal resolution is done using salience factors and Non-Pronominals using machine learning approach. Pronominal resolution refers to identification of a Noun phrase (NP) that is referred by a pronominal and Non-Pronominals are NP referring to another NP. In the next section we describe the system in detail.

2 System Description

In this section we give a detailed description of our system. The task is divided into two sub-tasks. They are

i) Pronominal resolution
ii) Non-pronominal resolution

2.1 Pronominal Resolution

Here we have identified salience factors and assigned weights for each factor. Before resolving the pronouns we identify whether a given pronoun is anaphoric or not. In example, (1) below, the pronoun "It", does not refer to any entity, and it is a pleonastic "it".

(1) "It will rain today"

In identifying the non-anaphoric pronouns such as "it" we use a CRFs engine, a machine learning approach. We build a language model using the above ML method to identify the non-anaphoric pronouns and the features used in training are word and it's POS in a window of five (two preceding and two following words to the pronoun). After the non-anaphoric pronoun identification, we resolve the anaphoric pronouns using a pronominal resolution system. Though we use salience factors based on the Lappin and Leass (1994), we have substantially deviated from the basic algorithm and have also used factors from Sobha (2008), where named entity and ontology are considered for resolution.

For identifying an antecedent for a pronoun we consider all the noun phrases before the pronoun in

Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 93–96,
Portland, Oregon, 23-24 June 2011. ©2011 Association for Computational Linguistics

the current sentence and in the four sentences preceding the current sentence. Those noun phrases which agree in PNG with the pronoun are considered as the possible candidates. The PNG is obtained using the gender data work of Shane Bergsma and Dekang Lin (2006). The possible candidates are scored based on the salience factors and ranked. The salience factors considered here are presented in the table 1.

Salience Factors	Weights
Current Sentence (sentence in which pronoun occurs)	100
For the preceding sentences up to four sentences from the current sentence	Reduce sentence score by 10
Current Clause (clause in which pronoun occurs)	100 – for possessive pronoun 50 – for non-possessive pronouns
Immediate Clause (clause preceding or following the current clause)	50 – for possessive pronoun 100 – for non-possessive pronouns
Non-immediate Clause (neither the current or immediate clause)	50
Possessive NP	65
Existential NP	70
Subject	80
Direct Object	50
Indirect Object	40
Compliment of PP	30

Table 1: Salience Factors and weights

Improving pronominal resolution Using Name Entity (NE) and WordNet: Pronouns such as “He”, “She”, “I” and “You” can take antecedents which are animate and particularly having the NE tag PERSON. Similarly the pronoun “It” can never take an animate as the antecedent. From the WordNet we obtain the information of noun category such as “person”, “object”, “artifact”, “location” etc. Using the NE information provided in the document and the category information in WordNet, the irrelevant candidates are filtered out from the possible candidates. Thus the antecedent and pronoun category agrees.

The highest ranked candidate is considered as the antecedent for the particular pronoun.

In TC and BC genres, the pronouns “I” and “you” refer to the speakers involved in the conversation. For these pronouns we identify the antecedent using heuristic rules making use of the speaker information provided.

2.2 Non-pronominal Coreference resolution

In identifying the Non-pronominal as said earlier, we have used a CRFs based machine learning approach. CRFs are well known for label sequencing tasks such as Chunking, Named Entity tagging (Lafferty et al, 2001; Taku Kudo 2005). Here we have CRFs for classification task, by using only the current state features and not the features related to state transition. The features used for training are based on Soon et al (2001). We have changed the method of deriving, values of the features such as String match, alias, from the Soon el al method and found that our method is giving more result. The features used in our work are as follows.

a) Distance feature – same as in Soon et al

b) Definite NP - same as in Soon et al

c) Demonstrative NP – same as in Soon et al

d) String match – (Not as Soon et al)the possible values are between 0 and 1. This is calculated as ratio of the number of words matched between the NPs and the total number of words of the anaphor NP. Here we consider the NP on the left side as antecedent NP and NP on the right side as anaphor NP.

e) Number Agreement – We use the gender data file (Bergsma and Lin, 2006) and also the POS information

f) Gender agreement – We use the gender data file (Bergsma and Lin, 2006)

g) Alias feature – (Not as in Soon et al) the alias feature takes the value 0 or 1. This is obtained using three methods,

i) Comparing the head of the NPs, if both are same then scored as 1

ii) If both the NPs start with NNP or NNPS POS tags, and if they are same then scored as 1

iii) Looks for Acronym match, if one is an acronym of other it is scored as 1

h) Both proper NPs – same as Soon et al.

i) NE tag information.

The semantic class information (noun category) obtained from the WordNet is used for the filtering purpose. The pairs which do not have semantic feature match are filtered out. We have not used the appositive feature described in Soon et al (2001), since we are not considering appositives for the coreference chains.

The feature template for CRF is defined in such a way that more importance is given to the features such as the string match, gender agreement and alias feature. The data for training is prepared by taking all NPs between an anaphor and antecedent as negative NPs and the antecedent and anaphor as positive NP.

The core CRFs engine for Non-pronominal resolution system identifies the coreferring pairs of NPs. The Coreferring pairs obtained from pronominal resolution system and Non-pronominal system are merged to generate the complete coreference chains. The merging is done as follows: A member of a coreference pair is compared with all the members of the coreference pairs identified and if it occurs in anyone of the pair, then the two pairs are grouped. This process is done for all the members of the identified pairs and the members in each group are aligned based on their position in the document to form the chain.

3 Evaluation

In this section we present the evaluation of the complete system, which was developed under the closed task, along with the independent evaluation of the two sub-modules.

a) Non-anaphoric detection modules
b) Pronominal resolution module

The data used for training as well as testing was provided CoNLL-2001 shared task (Pradhan et al., 2011), (Pradhan et al., 2007) organizers. The results shown in this paper were obtained for the development data.

The non-anaphoric pronoun detection module is trained using the training data. This module was evaluated using the 91files development data. The training data contained 1326 non-anaphoric pronouns. The development data used for evaluation had 160 non-anaphoric pronouns. The table 2 shows the evaluation, of the non-anaphoric pronoun detection module.

The Pronominal resolution module was also evaluated on the development data. The filtering of non-anaphoric pronouns helped in the increase in precision of the pronoun resolution module. The table 3 shows the evaluation of pronoun resolution module on the development data. Here we show the results without the non-anaphor detection and with non-anaphor detection.

Type of pronoun	Actual (gold standard)	System identified Correctly	Accuracy (%)
Anaphoric Pronouns	939	908	96.6
Non-anaphoric pronouns	160	81	50.6
Total	1099	989	89.9

Table 2: Evaluation of Non-anaphoric pronoun

System type	Total Anaphoric Pronouns	System identified pronouns	System correctly Resolved Pronouns	Precision (%)
Without non-anaphoric pronoun detection	939	1099	693	63.1
With non-anaphoric pronoun detection	939	987	693	70.2

Table 3: Evaluation of Pronominal resolution module

The output of the Non-pronominal resolution module, merged with the output of the pronominal resolution module and it was evaluated using scorer program of the CoNLL-2011. The evaluation was done on the development data, shown in the table 4.

On analysis of the output we found mainly three types of errors. They are

a) Newly invented chains – The system identifies new chains that are not found in the gold standard annotation. This reduces the precision of the system. This is because of the string match as one of the features.

Metric	Mention Detection			Coreference Resolution		
	Rec	Prec	F1	Rec	Prec	F1
MUC	68.1	61.5	64.6	52.1	49.9	50.9
BCUBED	68.1	61.5	64.6	66.6	67.6	67.1
CEAFE	68.1	61.5	64.6	42.8	44.9	43.8
Avg	68.1	61.5	64.6	53.8	54.1	53.9

Table 4: Evaluation of the Complete System

b) Only head nouns in the chain – We observed that system while selecting pair for identifying coreference, the pair has only the head noun instead of the full phrase. In the phrase “the letters sent in recent days”, the system identifies “the letters” instead of the whole phrase. This affects both the precision and recall of the system.

c) Incorrect merging of chains – The output chains obtained from the pronominal resolution system and the non-pronominal resolution system are merged to form a complete chain. When the antecedents in the pronominal chain are merged with the non-pronominal chains, certain chains are wrongly merged into single chain. For example “the chairman of the committee” is identified as coreferring with another similar phrase “the chairman of executive board” by the non-pronominal resolution task. Both of these are actually not referring to the same person. This happens because of string similarity feature of the non-pronominal resolution. This merging leads to building a wrong chain. Hence this affects the precision and recall of the system.

4 Conclusion

We have presented a coreference resolution system which combines the pronominal resolution using refined salience based approach with non-pronominal resolution using CRFs, machine learning approach. In the pronominal resolution, initially we identify the non-anaphoric pronouns using CRFs based technique. This helps in improving the precision. In non-pronominal resolution algorithm, the string match feature is an effective feature in identifying coreference. But, this feature is found to introduce errors. We need to add additional contextual and semantic feature to reduce above said errors. The results on the development set are encouraging.

References

Shane Bergsma, and Dekang Lin. 2006. Bootstrapping Path-Based Pronoun Resolution. In Proceedings of the Conference on Computational Lingustics / Association for Computational Linguistics (COLING/ACL-06), Sydney, Australia, July 17-21, 2006.

John Lafferty, Andrew McCallum, Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001). 282-289.

S. Lappin and H. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. Computational Linguistics, 20(4):535–562, 1994.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, Nianwen Xue. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011).

Sameer Pradhan and Lance Ramshaw and Ralph Weischedel and Jessica MacBride and Linnea Micciulla. 2007. Unrestricted Coreference: Identifying Entities and Events in OntoNotes. In Proceedings of the IEEE International Conference on Semantic Computing (ICSC)". Irvine, CA, September 17-19, 2007.

Sobha, L. 2008. Anaphora Resolution Using Named Entity and Ontology. In Proceedings of the Second Workshop on Anaphora Resolution (WAR II), Ed Christer Johansson, NEALT Proceedings Series, Vol. 2 (2008) Estonia. 91-96.

W. M. Soon, H. T. Ng, and D. C. Y. Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics, 27(4):521–544.

Taku Kudo. 2005. CRF++, an open source toolkit for CRF, http://crfpp.sourceforge.net .