Survey on Coreference Resolution, Relation and Event Extraction

Devendra Kumar Verma and Pushpak Bhattacharyya
Dept. of Computer Science and Engineering
Indian Institute of Technology, Bombay
Mumbai, Maharashtra, India
{devendra2689,pushpakbh}@gmail.com

ABSTRACT
Information Extraction refers to the automatic extraction of structured information such as relationships between entities and event and its arguments from unstructured sources. Information extraction is a branch of natural language processing that has a wide range of applications, including question answering, knowledge base population, information retrieval etc. The extraction of structure from noisy, unstructured sources is a challenging task.

INTRODUCTION
Early systems were rule-based with manually coded rules [4, 14, 20]. As manual coding of rules became tedious, algorithms for automatically learning rules from examples were developed [2, 6, 9, 21]. As extraction systems were targeted on more noisy unstructured sources, rules were found to be too brittle. Then came the age of statistical learning, where in parallel two kinds of techniques were deployed: generative models based on Hidden Markov Models [1] and conditional models based on maximum entropy [5, 12]. Both were superseded by global conditional models, popularly called Conditional Random Fields [13]. As the scope of extraction systems widened to require a more holistic analysis of a document’s structure, techniques from grammar construction were developed. In spite of this journey of varied techniques, there is no clear winner. Rule-based methods and statistical methods continue to be used in parallel depending on the nature of the extraction task. There also exist hybrid models that attempt to reap the benefits of both statistical and rule-based methods. Deep learning is the new state-of-the-art paradigm which is widely used in most of the information extraction systems. We will be covering co-reference resolution, relation extraction, event and its arguments extraction in details here.

COREREFERENCE RESOLUTION
Coreference resolution has been researched for many years, where the initial approaches were mostly knowledge-driven rule based approaches. These approaches continued to dominate till the advent of data-driven approaches. Along with discussing the existing approaches in rule-based and data-driven paradigms, we will also be discussing the linguistic aspects of coreference resolution.

Linguistic and Other Considerations
Linguistic factors such as syntactic constraints, semantic cues like gender, world knowledge and knowledge of textual structure are of primary importance to resolution of coreference (Crawley et al., 1990). Most of the widely discussed features which are found effective for the task, are motivated by these factors. Interpretation of noun phrases in many cases depend on the linguistic context, considering the discourse situation (Bean and Riloff, 2004). Syntactic structure and syntactic preferences play a major role. Majority of the pronouns and their antecedents occur in the subject position of the sentence. Kertz et al. (2006) discusses parallel function preference stating that an anaphora and its antecedent tend to have the same grammatical role. An anaphoric mention with a subject grammatical role is likely to have an antecedent with subject role.

Rule Based Approaches
Most of the initial approaches in coreference resolution were highly linguistic oriented till the introduction of data-driven approaches. Winograd (1972) proposed a coreference resolution system as part of an automated English understanding system, considering all preceding noun phrase candidates for probable antecedent and rate them based on their syntactic position. Hobbs (1978) discusses one of the earliest syntactic approaches; Hobbs algorithm. For a pronoun mention, this algorithm utilizes constituency parse tree to identify the antecedent. This method also incorporates syntactic constraints and semantic considerations through rules. Rich and Luper-Foy (1988) scores the probable antecedents after evaluating each antecedent with a set of defined constraint sources. The final score of an antecedent candidate is a function of score given by each constraint source and the confidence associated with the constraint score.

Even after the introduction of data driven approaches in coreference resolution, there were a few rule-based systems exhibiting matching performance compared to the state-of-the-art systems. Among these, the prominent one is the Stanford coreference resolution system (Raghunathan et al., 2010)
based on a multi-sieve rule-based approach. This applies
deterministic coreference models at different phases in the de-
sceding order of their strength in deciding coreference. The ini-
ital passes resolve exact match, appositives, relative pron-
oun etc. The last pass is dedicated to resolve pronouns.

Data-driven Approaches

Features for Coreference Features provide the essential
cues for checking coreferent relation between mentions in any machine learning based approach. These features are
strongly motivated by the linguistic clues for coreference. Mos-
ly the features are computed taking two mentions at a
time; except for the approaches where the belongingness of a
mention to a cluster of mentions is evaluated. Features are
generally classified as lexical, grammatical (NP type, NP
property/relationship, syntactic pattern), semantic and posi-
tional.

Apart from these features, different coreference reso-
lution systems have proved the usefulness of various other features. Stamborg et al. (2012) discusses incorporating lin-
guistic phenomena and discourse properties to the features. They discuss some novel features including discourse and
type of document. For coreference resolution in certain lan-
guages (eg. Spanish), feature to check if a mention is an ellipt-
tical pronoun is crucial (Recasens and Hovy, 2009). Rosiger
and Riester (2015) discusses prosodic features for resolving
coreference in spoken text.

Modeling Coreference Resolution: Since the decision of
coreference involves many mentions in a text, there are dif-
ferent ways the problem can be modeled. Mostly in all these
methods, at the root level the comparison is between the
two mentions at a time. There were different attempts to
model the problem of coreference resolution. Some existing
approaches experiment with different coreference models to
demonstrate the impact of their contribution.

Mention-Pair Model: Mention-pair model has a classi-
fication step followed by clustering. Classification take into
consideration two mentions at a time, classifying them as coreferent or not. For an anaphoric mention $m_k$, the classify-
ing step checks if a candidate antecedent $m_i$ is coreferent.
Features are computed for each mention pair. In a supervised
approach a training instance is created with a mention and its
closest antecedent, and for the same anaphoric mention neg-
ative instances are created by pairing with mentions occur-
rning before it and after its closest antecedent. During testing,
the clustering step following the classification, identifies the
best antecedent for an anaphoric mention. Clustering picks
the best antecedent from the candidate antecedents which are
identified coreferent with the anaphoric mention after classi-
fication. This forms distinct coreferent chains in a document.

Entity-Mention Model: When classification confines to a
mention pair in the mentionpair model, entity-mention
model compares with previously identified partial clusters.
The classifier determines if a mention belongs to one among
the partial clusters occurring before this mention. Each train-
ing instance contains a mention and a cluster and the com-
puted features include cluster level features and features per-
taining to the mention under consideration.

The problem of coreference resolution specifically for
Legal domain has received relatively limited attention in lit-
erature. The literature broadly categorized into two streams.
One focuses on anaphora resolution [3] and the other ad-
dresses the problem of Named Entity Linking. Anaphora
Resolution is a sub-task of Coreference Resolution where
the focus is to find an appropriate antecedent noun phrase
for each pronoun. The task of Named Entity linking [7, 8,
10] focuses on linking the names of persons / organizations
and Legal concepts to corresponding entries in some exter-
nal database (e.g. Wikipedia, Yago). In comparison, our
approach focuses on grouping all the coreferent mentions to-
gether including generic NPs.

Even in the general domain, the problem of coreference
resolution remains an open and challenging problem [15].
Recently, Peng et al. [18, 19] have proposed the notion of
Predicate Schemas and used Integer Linear Programming for
coreference resolution. In terms of problem definition and
scope, our work is closest to them as they also focus on all
three types of mentions, i.e. named entities, pronouns and
generic NPs.

Supervised Approaches

Supervised approaches gained popularity by mid-1990 in re-
solving coreference. Based on the aforementioned ways to
model the task of coreference resolution, the machine learn-
ing approaches can be broadly classified into two; one is a
2 step approach with a binary classification followed by
clustering and the second is a ranking approach (Zheng et
al., 2011). One of the earlier statistical approaches is by Da-
gan and Itai (1990), where word co-occurrences are taken
into account to disambiguate pronouns, but restricted to the
pronoun it. For an anaphoric mention as subject of a verb, the men-
tion among the candidate antecedents having maximum co-
occurrences with the same verb as subject is selected as the
antecedent. In another different attempt by Ge et al. (1998)
proposed a statistical framework for resolution of third person
pronouns which learns a probabilistic model using Penn Wall
Street journal Tree-bank (Riezler et al., 2002). For a can-
didate antecedent to an anaphoric mention, this computes prob-
ability values for certain factors (eg. distance, co-occurrence
patterns etc.) based on the probability values computed over
the training data. These probabilities are multiplied to com-
pute the probability associated with a candidate antecedent.

Introducing mention-pair model, one of the widely used
modeling paradigm for coreference resolution, Aone and
Bennett (1995) introduced a coreference resolution system
for Japanese coreference resolution. They experimented C4.5
decision tree classifier for mention pair classification. Dif-
ferent variants of their approach are evaluated against their
own previ-ously designed solver based on manually selected
knowledge sources. During the same time, McCarthy and
Lehner (1995) and Connolly et al. (1997) came up with
a machine learning based approach for English. Following
the same modeling paradigm of Aone and Bennett (1995),
Soon et al. (2001) built a machine learning based coreference resolution system focusing more on the design of features. They employed C5 decision tree algorithm for mention pair classification, and the system gives matching performance with the then existing rule-based systems on MUC-6 and MUC-7 datasets. Extending this work, Ng and Cardie (2002b) introduced a deeper set of features for coreference. Ng and Cardie (2002a) modified this approach by determining the anaphoricity of a noun phrase as a pre-processing step. Bergsma et al. (2008) determines the anaphoricity of a noun phrase through a method based on context distribution. Ram and Devi (2012) discusses a CRF based approach for determining anaphoricity, and Ng (2009) proposes a graph cut based anaphoricity determination algorithm. Yurykina (2006) experimented with different classifiers extending the feature set from the conventional set of features with more linguistically motivated features. There has been several attempts to improve the discussed methods through utilization of semantic knowledge from diverse sources. Along with introducing Ontonotes; the present widely used dataset for coreference resolution, Pradhan et al. (2007b) introduced a baseline model with classifier as Support Vector Machine.

RELATION EXTRACTION

End-to-end relation extraction refers to identifying boundaries of entity mentions, entity types of these mentions and appropriate semantic relation for each pair of mentions. Traditionally, separate predictive models were trained for each of these tasks and were used in a pipeline fashion where output of one model is fed as input to another. But it was observed that addressing some of these tasks jointly results in better performance.

Most of the past work in relation extraction deals with relations occurring within a sentence and having only two arguments. Open IE systems have achieved a notable measure of success on massive, open-domain corpora drawn from the Web, Wikipedia, and elsewhere. (Banko et al., 2007; Wu and Weld, 2010; Zhu et al., 2009). The output of Open IE systems has been used to support tasks like learning selectional preferences (Ritter et al., 2010), acquiring common sense knowledge (Lin et al., 2010), and recognizing entailment (Schoen- mackers et al., 2010; Berant et al., 2011). In addition, Open IE extractions have been mapped onto existing ontologies (Soderland et al., 2010). Anthony Fader et al. (2011) [11] imposed two constraints to identify relation phrases

1. **Syntactic Constraint**: It helps in identifying relation phrases expressed by a verb-noun combination by matching POS tag pattern.

2. **Lexical Constraint**: While matching using syntactic constraint there are many irrelevant information gets extracted to reduce the number of those this constraint enforced. According to this constraint a valid relation phrase should take many distinct arguments in a large corpus.

A tool REVERB is a novel open information extractor, based on the constraints defined above. REVERB first identifies relation phrases that satisfy the syntactic and lexical constraints, and then finds a pair of NP arguments for each identified relation phrase. The resulting extractions are then assigned a confidence score using a logistic regression classifier.

[Swampillai and Stevenson 2011] observed that the structured features which are generally used for sentence relation extraction can be easily adapted for inter-sentence relations. They proposed to introduce a dependency link between the root nodes of parse trees containing the given pair of entities and developed features based on the shortest path connecting the pair of entities in the new fused tree. [Quirk and Poon 2016] proposed a new approach for cross-sentence relation extraction using distant supervision. They proposed a document-level graph representation that incorporates both intra-sentential dependencies and inter-sentential relations such as adjacency and discourse relations. [Peng et al. 2017] proposed a general framework for N-ary cross-sentence relation extraction, based on graph long short-term memory networks. They use the same document graph as proposed by [Quirk and Poon 2016] and it acts as a backbone upon which a graph LSTM is constructed.

EVENT EXTRACTION

Early research on event extraction has primarily focused on local sentence-level representations in a pipelined architecture (Grishman et al., 2005; Ahn, 2006). After that, higher level features has been investigated to improve the performance (Ji and Grishman, 2008; Gupta and Ji, 2009; Patwardhan and Riloff, 2009; Liao and Grishman, 2010; Liao and Grishman, 2011; Hong et al., 2011; McCloskey et al., 2011; Huang and Riloff, 2012; Li et al., 2013). Besides, some recent research has proposed joint models for EE, including the methods based on Markov Logic Networks (Riedel et al., 2009; Poon and Vanderwende, 2010; Venugopal et al., 2014), structured perceptron (Li et al., 2013; Li et al., 2014b), and dual decomposition (Riedel et al. 2009; 2011a; 2011b)).

Nguyen and Grishman (2015b) study domain adaptation and event detection via CNNs while Chen et al. (2015) apply dynamic multi-pooling CNNs for EE in a pipelined framework. Nguyen and Grishman (2015b) [17] studied the event detection problem using convolutional neural networks (CNNs) that overcome the two fundamental limitations of the traditional feature-based approaches to this task: complicated feature engineering for rich feature sets and error propagation from the preceding stages which generate these features. The experimental results show that the CNNs outperform the best reported feature-based systems in the general setting as well as the domain adaptation setting without resorting to extensive external resources.

Nguyen and Grishman (2016)[16] proposed to do event extraction in a joint framework with bidirectional recurrent neural networks, thereby benefitting from the advantages of the two models as well as addressing issues inherent in the existing approaches. They systematically investigate different memory features for the joint model and demonstrate that the proposed model achieves the state-of-the-art performance on the ACE 2005 dataset.
CONCLUSION

Coreference Resolution, Relation Extraction and Event and its Extraction are some of the challenging tasks of NLP. In this survey paper we have covered various techniques for coreference resolution, which includes linguistic consideration, rule based approaches, data driven approaches. We have also discussed about the coreference resolution modeling techniques as well. For relation extraction we discussed open domain relation extraction. For event various techniques including deep learning based models like CNN and RNN were discussed.

REFERENCES


