IEEE Symposium: Computational Intelligence for Security and Defence Applications, Ottawa, Canada, July 2009. Copyright IEEE.

He says, she says. Pat says, Tricia says. How much reference resolution matters for entity extraction, relation extraction, and social network analysis

Jana Diesner, Kathleen M. Carley

Anaphora resolution (AR) identifies the entities that pronouns refer to. Coreference resolution (CR) associates the various instances of an entity with each other. Given our data, our findings suggest that deduplicating and normalizing text data by using AR and CR impacts the literal mention, frequency, identity, and existence of about 75% of the entities in texts. Results are more moderate on the relation level: 13% of the links are modified and 8% are removed. Performing social network analysis on the relations extracted from texts leads to findings contrary to the results from corpus statistics: AR and CR cause different directions in the change of network analytical measures, AR alters these measures more strongly than CR does, and each technique identifies a different set of most crucial nodes. Bringing the results from corpus statistics and social network analysis together suggests that CR is more effective in normalizing entities, while AR is a more powerful technique for splitting up generic nodes into named entities with adjusted weights. Data changes due to AR and CR are qualitatively and quantitatively meaningful: the statistical properties of entities and relations change along with their identities. Consequently, the relational data represent the underlying social structure more truthfully. Our results can support analysts in eliminating some misinterpretations of graphs distilled from texts and in selected those nodes from social networks on which reference resolution should be performed.

I. Introduction

The analysis of natural language text data often involves the identification of the relevant pieces of information that are needed for answering a question or solving a problem. An integral part of this process is the application of various natural language processing (NLP) techniques. In order to apply NLP techniques in an informed and reliable fashion and to support the drawing of reasonable conclusions and meaningful inferences from text analysis results, a precise understanding of the potential impact of NLP techniques on the data is needed. For example, our prior work has shown that slight differences in the choices that one has to make when implementing a parts of speech (POS) tagger, such as removing noisy symbols that are neither letters nor numbers prior to model training or not, can lead to significantly different POS tagging accuracy rates [1]. In this paper we investigate the impact of reference resolution on relation extraction and social network data. Reference resolution is a commonly used NLP techniques that identifies the referent that a referring expression refers to [2, 3]. Relation extraction distills relational data from texts, such as who is located where and communicates with whom about what. These data are often referred to as semantic networks [4-6]. If the data represent interactions between social and socio-technical agents they are also called social networks [7-9]. Social networks can be examined by using social network analysis methods [10]. At a minimum, relation extraction involves the identification of entities - a process also known as entity extraction - and of the relations between them [11]. Thus, entity extraction is a precondition for relation extraction. Advanced relation extraction approaches support the classification of entities and/ or relations according to pre-defined or data-induced categorization schemas [12, 13]. In graph representations of relational data, which are to input for social network analysis, entities are called nodes and relations are links or edges.

In this study, we do not ask how accurate reference resolution techniques are, but how much of a difference their usage makes for relation extraction and subsequent network analyses. Why is this study relevant? State of the art entity and relation extraction approaches apply probabilistic and machine learning techniques to large volumes of text data (see e.g. [12, 14, 15]). These techniques exploit information provided by a plethora of various NLP subroutines; with reference resolution being one of them. The resulting data are often too voluminous and complex to be verified by humans. Knowing the precise impact of each involved sub-routine increases the control over multi-stage analytical processes and the understanding of the results. Our findings can support researchers and practitioners from different areas such as social network analysis, (computational) linguistics, and intelligence in deciding how much effort to spend on reference resolution in projects involving relation extraction and in interpreting respective network analysis results.

This work was supported by the following grants: National Science Foundation (NSF) IGERT 9972762, ONR MMV N00014-06-1-0104, ARL Alion STM1193391, ARI W91WAW07C0063, Vibes 1193343TO102, and AFOSR MURI FA9550-05-1-0388. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the Army Research Lab, the Army Research Institute, the Office of Naval Research, the Department of Defense, or the U.S. government.

Jana Diesner (corresponding author) and Kathleen M. Carley are both with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA (corresponding author: phone: 412-268-3163; fax: 412-268-1744; e-mail: jdiesner@andrew.cmu.edu; second author: e-mail: kathleen.carley@cs.cmu.edu).

This report is structured as follows: The background section familiarizes the reader with the reference resolution and relational data analysis techniques addressed herein. The method section describes the dataset that we use and reports on the experimental design that we employed in order to test our hypotheses. In section four we present the results. The article concludes with a discussion of the relevance, implications, and limitations of our findings.

II. Background

Reference resolution (RR) involves anaphora and coreference resolution. Anaphora resolution (AR) identifies the antecedent *A* that an anaphoric expression - short anaphor - *B* refers to [2]. Typically, *A* is a noun phrase and precedes *B*, which usually is a pronoun, in the text. *A* is only considered to be an antecedent of *B* if *A* is required for resolving *B*. Thus, the relationship between *A* and *B* is non-symmetric, non-reflexive, and non-transitive [16]. Coreference resolution (CR) identifies the set of entities that refer to the same referent *C* [3]. These entities are typically noun phrases. Entity *C* may or may not be explicitly mentioned in the text. *A* and *B* are only said to co-refer to each other if they both unambiguously represent *C*, such that *A*=*C* and *B*=*C*. Therefore, coreferences are symmetric, reflexive, and transitive equivalence relationships [16].

How do AR and CR relate to each other? If an anaphor *B* and its antecedent *A* refer to the same entity, *A* and *B* are coreferential. However, there is no deterministic or set-theoretic relationship between AR and CR, i.e. an anaphoric and a coreferential relation may overlap, but not all cases of AR are also cases of CR and vice versa. Another difference between AR and CR is that for resolving a given *B*, in AR, *A* has to be interpreted within the context of the text in which both phrases occur, while in CR, interpreting *A* is not required for testing which entity *C* a *B* is identical to. For example, in the phrase “the 2000 democratic presidential candidate and 2007 Nobel Peace Prize winner”, both mentions of a person refer to the real-world entity *C* = Al Gore, but an interpretation of entity *A* (candidate) is not required for resolving entity *B* (winner). In contrast to that, resolving the referential expression *B* = “he” in the phrase “Al Gore ran for president in 2000. In 2007, he won the Nobel Peace Prize”, with Al Gore being the antecedent *A*, requires an interpretation of the text preceding *B*.

AR and CR are active research topics in NLP. Respective work focuses on improving the accuracy of automated RR techniques. The best accuracy rates reported, which strongly depend on the applied resolution method, data set, and evaluation metrics, achieve a harmonic mean between precision and recall of over 80% for AR, and of about up to 70% for CR [17-20]. The top scoring techniques are often based on supervised machine learning.

Relation extraction involves three steps, which are typically performed in the following order [7, 9, 11]:

1. data preprocessing
2. entity identification
3. relation identification

These three processes are not independent of each other in the sense that decisions made in one stage impact results obtained in the next stage. Both, AR and CR, are normalization and deduplication techniques that can be applied as preprocessing steps for relation extraction: AR helps to associate pronouns with the entity or entities that the pronoun refers to, while CR maps multiple instances of an entity to a unique key identifier for that entity. Both routines potentially increase the number of mentions per unique entity. In network terms, the number of entity mentions are node weights and the number of relation mentions are edge or link weights. When AR and/ or CR are applied in the preprocessing stage, they might impact the identity, literal mention (i.e. spelling), and weight of entities and relations, and thereby also the results obtained from running network analysis on the extracted relational data. Since the magnitude of this impact is not well understood yet, we herein raise and address the following research question: How much variation in network structure and values of network analytic measures are due to the application of AR and/ or CR?

III. Method

In order to answer our research question we first need to precisely define the effects that AR and CR can have on text and network data. Let’s assume that prior to AR and CR, every entity is unique and has a weight of one. When performing AR, every anaphor that is not referred to by a name (unspecified entity) and that can be mapped to an anteceding referent that is referred to by a name (named entity) is associated with that named entity. When doing CR, all entities that refer to the same referent are consistently converted into one unique key identifier for that referent. Thereby, co-referring entities are associated with each other. While AR does not change the number of unique named entities, CR potentially reduces this number. Both, AR and CR can increase the weight or empiric evidence per unique entity. While AR mainly reduces the number of unspecified entities, CR only leads to this reductive effect if multiple unspecified instances of an entity, such as a set of unresolved pronouns, are identified to be co-referring to the same entity that is also unspecified. Table I summarizes the set of possible effects that are AR and CR can have on entity extraction.

TABLE I

Possible impacts of AR and CR on entity extraction

Case	Entity		Impact on technique		Impact on named entities	
	Name or Nominal	Pronoun	Anaphora Resolution	Coreference Resolution	Number	Weight (impacted entities)
1	N=1	0	irrelevant	irrelevant	none	none
2	0	N=1	failure	irrelevant	none	none
3	N>1	0	irrelevant	success	decrease	increase
4	0	N>1	irrelevant	success	none*	none**
5	N=1	N	success	irrelevant	none*	increase
6	N>1	N	success	success	decrease	increase

* Decrease of number of unspecified entities

** Increase of weight of impacted unique unspecified entities

The cells labeled as "success" in Table 1 represent the desired effects of applying RR: the various instances and mentions of unique entities, such as repetitions, spelling variations, abbreviations, and pronouns are gathered together and consistently associated with or converted into one key identifier per unique entity.

If one or both nodes in a relation are resolvable anaphors, the respective node names can be replaced with the according named entity. This process does not alter the link weight. If the nodes *A* and *B* in a link are coreferences of two nodes *C* and *D* in another link, such that *A=C* and *B=D*, the two links can be collapsed into one link while increased the link weight by one. If further links are mapped onto this link, the link weight is increased accordingly. AR and CR on the entity level are preconditions for impacts of RR on the relation extraction and network analysis level.

From these definitions and the logic presented in Table I we derive the following hypotheses:

1. AR and CR each increase the number of mentions per unique and named entity.
2. AR decreases the number of unspecified entities and relations.
3. CR decreases the number of unassociated (i.e. unrelated) named entities and relations.
4. Combining AR and CR is more effective in increasing the mentions per unique named entity and relations and in decreasing the number of unspecified and unassociated entities and relations than either technique alone.

In order to empirically test our hypotheses we need a sufficiently large corpus in which anaphors and coreferences are resolved with high reliability. We use version 1.0 of the ACE2 corpus (ACE in the following) for this purpose [21]. This corpus is part of the Automatic Content Extraction program, which is administered by the National Institute of Standards and Technology (NIST) and offers annual competitions on various NLP problems. Cutting-edge AR and CR techniques have been developed and validated by using ACE [22]. The ACE data set was created by trained human coders and is available through the Linguistic Data Consortium (LDC). The data contains 518 annotated files and was produced by various newspaper, newswire and broadcast agencies during the first half of 1998. ACE can be split into three subsets of data according to these genres, which we did for this study.

In ACE, anaphors are marked as pronouns, and also include terms like one, some and there. Phrases applicable to CR as well as potential resolutions for anaphors are annotated as names and nominals. In this study we refer to names and nominals as named entities, and to pronouns as unspecified entities. For this project, we do not resolve anaphors and coreferences algorithmically, but work with the resolutions provided by human coders. Thereby we resemble the gold standard, which assumes human judgment to be the correct solution [23]. This strategy allows us to make non-probabilistic statements about the impact of AR and CR, because the correct solutions for either routine are explicitly given in the data and were previously validated by trained humans.

Current RR algorithms achieve accuracy rates of less than 100%, and no algorithm might ever return perfectly correct reference resolution results. Our results are based on the judgment of trained people who aimed to deliver the best RR results that humans can possibly provide. Therefore, our findings report on the upper bound of the impact of highly accurate AR and CR on entity and relation extraction and network analysis. However, the findings about the influence of RR on relation extraction and network analysis strongly depend on the chosen relation formation procedure. The various methods that have been developed for this purpose use lexical, syntactical [24, 25], semantic [6], logical [6, 9], and proximal [7, 26] information from texts. The relation identification performed by human coders for ACE operates on the sentence level. In addition, if the instances (*C*, *D*) of a pair of nodes *A* and *B*, such that *A=C* and *B=D,* are identified to form the same type of relationship in a text, i.e. the link from *A* to *B* is of the same type as the link from *C* to *D*, the respective relation is annotated to have multiple mentions (in this case two). If *A=C* and *B=D*, but the link from *A* to *C* represents a different type of relationships than the link from *B* to *D*, these relationships are marked as different relations in ACE. Finally, the data contained 20 fully redundant relations (same type of relationship between identical nodes at the same text position), which we deduplicated.

IV. Results

We start this section with a quantitative description of the data: the newswire, newspaper, and broadcast data each account for roughly a third of all entities and entity mentions (Fig. 1). The accounts of written language (newswire and newspaper) are fairly similar to each other regarding the distribution of node types (pronoun, name, nominal) and differ in that respect from the transcripts of spoken language (broadcast). The broadcast data has proportionally more pronouns, less names and nominals, and less mentions per entity than the other two subsets. Across all datasets, the set of unique entities is dominated by nominals, while additional references to entities are mainly realized through names. The recurrence of previously introduced concepts occurs most often in newspaper data, and least often in broadcast news.

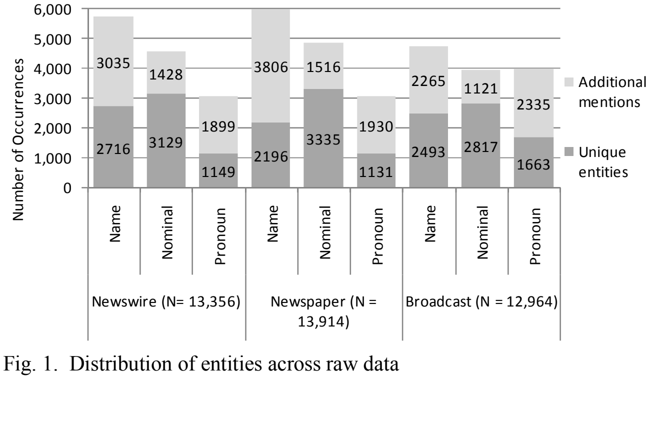

Fig. 1. Distribution of entities across raw data

Next, we present the results for the entity and relation level, which both are forms of corpus statistics, and then on the network level. More than 60% of all entity mentions are potentially affected by RR (Fig. 2): pronouns account for roughly a quarter of the entities in the data (light gray bar at the top of the stacked bars). They are subject to AR. Additional mentions of unique entities constitute another 38% of the data (middle section of the bars) and can be affected by CR. For newswire and newspaper data, CR can make a bigger difference than AR, and vice versa for broadcast data. The final goal with RR is to map the pronouns and additional mentions to the set of unique entities, thereby reducing the mass of unspecified and unassociated entities while increasing the weight of unique entities. Interaction effects between AR and CR can lead to further data normalization: AR might convert an unspecified entity into a named entity which is then further aggregated by CR.

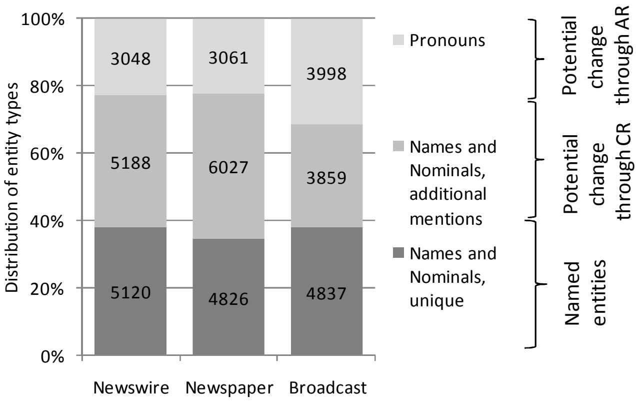

Fig. 2. Distribution of entity types across raw data

Using our methodology, anaphors are only irresolvable via AR if all mentions of the respective pronoun are pronouns themselves. The AR results (Fig. 3) show that for newswire and newspaper, more than 80% of all pronouns are resolvable. For broadcast, 67% of the pronouns can be mapped to a named entity. We speculate that for transcripts of spoken language, AR is complicated by the fact that these data have proportionally more pronouns to begin with, so that a smaller pool of names and nominals with which the pronouns can be associated is available. Our findings indicate that AR can serve as a powerful data normalization technique and confirm hypotheses one and two on the entity level.

For the newswire and newspaper data, more than 75% of the irresolvable pronouns are entities that are mentioned only once (single mentions). In the broadcast data, less than 66% of the pronouns that remain after AR are single mentions; the rest of them forms groups of multiple mentions per pronoun. These groups are subject to CR.

Our results for CR show that two thirds of all names and nominals are single mentions. They are applicable to CR (Fig. 4). The remaining one third of names and nominals are coreferenced. They carry about two thirds of the total weight of all named entities. These numbers suggest that CR can be a highly effective consolidation technique and confirm hypotheses one and three for the entity level.

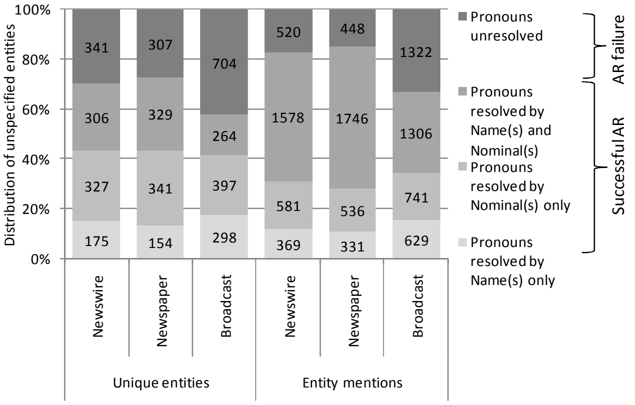

Fig. 3. Anaphora resolution results

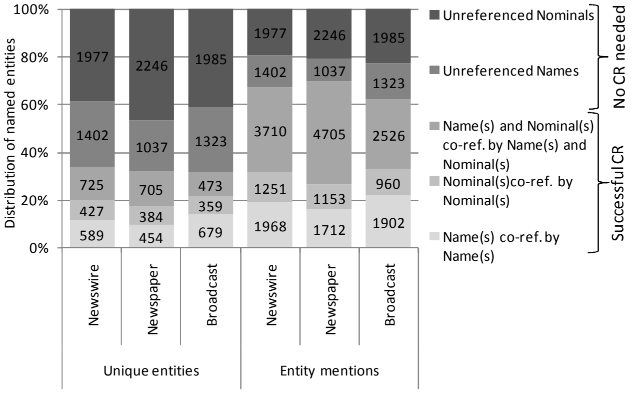

Fig. 4. Coreference resolution results

Putting the results for AR and CR together, as illustrated in Fig. 5., shows that only a quarter of all text terms identified as entities are not impacted by AR and/ or CR. These entities are either irresolvable anaphoric expression or entities that are mentioned only once.

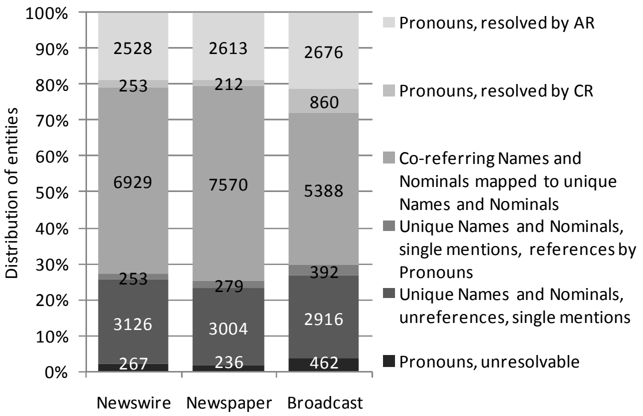

Fig. 5. Combined impact of AR and CR

Table II further details the separate and combined impact of AR and CR. These results show that CR contributes more strongly to the data normalization and consolidation effects than AR does. This finding partially reflects the distribution of entity classes across the data, and is furthermore aided by the fact that AR increases the set of named entities applicable to CR in the first place. Table II also shows that combining AR and CR clearly outperforms applying only either one technique in terms of decreasing the number of named and unspecified entities while increasing the weight of the affected unique entity. This result confirms hypothesis four on the entity level. Moreover, it illustrates how informed and reliable reference resolution helps to eliminate redundant data points while enriching the amount of information available on truly distinct entities. The numbering of the cases in Table II is analogous to the logic presented in Table I. The baseline for comparison in Table I is the raw data, in which the weight of each entity equals one. Deploying AR plus CR increases the weight of the impacted unique entities to almost five on average. However, the actual weights per entity that are higher than one vary widely and do not follow a normal distribution. The last four columns in Table II illustrate this point more precisely: Deploying both mechanisms in one round of preprocessing makes 38% of the unique entities carrying 75% of the total weight of entities, while the remaining 62% of unique entities only account for 25% of the weight.

Another interesting finding shown in Table II is case 4, which suggests that coreference resolution on unspecified entities which could not be resolved via AR has a minor yet meaningful impact on the data.

Table II

Impact of AR and/ or CR on number and weight of unique entities

Case	Impact on data					
	Decrease in number of entities	Average weight of impacted unique entities	Ratio of unique entities impacted	Ratio of total weight carried by impacted unique entities	Ratio of entities not impacted (weight of each = 1)	Ratio of total weight carried by not impacted entities
3 (CR)	-37.72%	4.13	19.3%	49.8%	80.7%	50.2%
4 (CR on pronouns)	-2.35%	3.42	1.0%	3.3%	99.0%	96.7%
5 (AR)	-19.56%	4.01	8.1%	26.0%	91.9%	74.0%
6 (AR and CR)	-59.63%	4.89	38.0%	74.9%	62.0%	25.1%

Transforming the number of all unique entities and their respective weights into log space provides a compact view on the skewed distribution of data and the magnitude of the investigated techniques' impact (Fig. 6): the higher a point is on the y-axis, the more mentions are associated with this entity. The curves intersect with the x-axis where entity weights start to equal one; thus representing single mention entities. The dotted line along the x-axis represents the raw data. Such a distribution is likely to be obtained in text analysis projects in which no effort to resolve references is made. The line for the combined effect of AR and CR has the largest area under the curve; indicating that higher weights for more entities were obtained than with any other strategy. This curve hits the x-axis last, which illustrates the fact that the least amount of single mention entities remains in the data to which AR plus CR were applied.

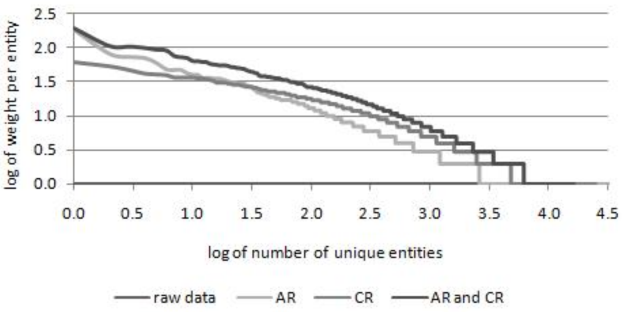

Fig. 6. Impact of AR and/ or CR on number and weight of entities

Not all entities are linked into relations. This can due to two reasons: First, some entities are really not related to any other entity (isolates). Second, in ACE, relation identification was performed within but not across sentences. Consequently, relations between entities that do not co-occur in the same sentence are not annotated. Both reasons add to the sparseness of relations across texts. In ACE, a third of the entity mentions and a little more than half of the unique entities are involved in relations formation. On average, 16% of the relations contain at least one pronoun (Table III). Depending on the dataset, in 76% to 87% of these relations, the pronouns can be resolved to named entities. Consequently, the identity and spelling of nodes in 11.3% to 14.5% of the relations are adjusted. This confirms hypothesis two on the network level.

Between 4.2% and 7.5% of relations that are composed of names and nominals only can be mapped to existing unique relations via CR (Table III). This confirms hypothesis three on the node level. Only three links that consist of pronouns only and for which both pronouns were irresolvable could be mapped to co-referring pronoun-only links. Since these effects are marginal we disregard them from the analysis on the relation level.

Combining AR and CR has a stronger impact on merging multiple instances of relations onto unique relations whose weights are increased accordingly than using either technique alone (last row in Table III). This confirms hypothesis four on the network level. While most of the relations reduction is due to CR, the majority of node spelling adjustments can be attributed to AR.

Table III

Impact of AR and/ or CR on number and weight of relations

Technique applied	Measure	Newswire	Newspaper	Broadcast
none (raw data)	number of relations	2884	2956	2267
AR	percentage of relations with pronoun	14.8%	16.7%	16.5%
	perc. of rel. with pronoun on which AR was successful	76.6%	87.0%	76.1%
	percentage of relations changed	11.3%	14.5%	12.5%
CR	perc. of relations with names and nominals only	85.2%	83.3%	83.5%
	relation reduction rate	4.2%	4.7%	7.5%
AR and CR	relation reduction rate	6.5%	7.9%	10.6%

Overall, the link normalization and deduplication effects due to RR are less strong on the relation level than on the entity level (compare Tables II and IV). While the average weight of unique entities impacted by AR plus CR increases from 1.0 to 4.9, the weight of unique relations affected by AR plus CR is only raised from 1.0 to 2.3. On the entity level, the weight of 38% of the unique is increased and cumulatively accounts for three quarters of the total entity weight, while on the relation level, the 17% of the unique relations whose weight is increased due to applying AR plus CR carry less than a quarter of the total weight.

Table IV
Impact of AR and/ or CR on number and weight of unique relations

Case	Impact on data					
	Relation reduction rate	Average weight of impacted relations	Ratio of unique relations impacted	Ratio of total weight carried by impacted relations	Ratio of relations not impacted (weight of each = 1)	Ratio of total weight carried by not imp. relations
CR	5.33%	2.15	4.9%	10.0%	95.1%	90.0%
AR	n.a.	n.a.	12.8%	12.8%	87.2%	87.2%
AR and CR	8.17%	2.25	17.4%	24.2%	82.6%	75.8%

In order to investigate the impact of reference resolution on social network analysis (SNA) we need to go one step further and perform SNA on the data resulting from relation extraction. We use the ORA software for this purpose [27]. While the previous analyses were based on unambiguous, explicit, and numeric identifiers for each entity and relation as specified in XML files that represent the ACE data, ORA uses the actual node names as input. Therefore, link aggregation is now not only based on matching index numbers, but in addition to that also on the matching spelling of nodes: when ORA encounters a node or an edge with the exact same spelling as a previously registered element it does not add another node or edge to its data registry, but increases the weight of the previously encountered element accordingly. This is common procedure in many SNA tools and libraries. To give an example, if the phrase “President Clinton” occurs in several documents it would always be mapped to the same unique node. If, on the other hand, some unspecified entity referred to as “he” is found in multiple files and cannot be resolved, all instances of this node are collapsed into one collective node labeled “he”, regardless of whether “he” refers to different social entities or not. If RR is performed not on a per document basis, as done in ACE, but on the corpus level, we suggest keeping identically spelled nodes that were identified to represent truly distinct entities separate in order to not to dilute the effects achieved by using AR and CR.

Table V presents the values of a set of key measures that are frequently used in SNA in dependence of the data preprocessing techniques investigated herein (for details on the measures see [27, 28]). The last three columns in Table V show the change from raw data to AR and CR applied alone and in combination. For the vast majority of these SNA measures, AR and CR exhibit opposite effects with respect to increasing or decreasing the value of a measure.

Table V
Value of SNA measure per routine and change in measure from raw data to applied routine

SNA measure	raw	AR	CR	ARCR	raw to AR	raw to CR	raw to AR+CR
Number of nodes	4048	4103	4002	4042	1.4%	-1.1%	-0.1%
Link Count	6910	6981	6716	6684	1.0%	-2.8%	-3.3%
Density	0.000	0.000	0.000	0.000	0.0%	0.0%	0.0%
Network (Ctrz.)-Betweenness	0.032	0.036	0.031	0.039	12.0%	-2.2%	23.0%
Network Ctrz.-Closeness	0.000	0.000	0.000	0.000	0.0%	0.0%	0.0%
Network Ctrz.-Total Degree	0.036	0.023	0.037	0.023	-37.3%	1.1%	-37.0%
Centrality-Bonacich Power (Av.)	1.996	1.967	2.019	2.002	-1.4%	1.2%	0.3%
Centrality-Eigenvector (Av.)	0.123	0.118	0.126	0.123	-3.8%	3.0%	0.2%
Diameter	4048	4103	4002	4042	1.4%	-1.1%	-0.1%
Diffusion	0.109	0.123	0.107	0.114	13.1%	-1.7%	4.1%
Fragmentation	0.220	0.215	0.227	0.224	-2.2%	3.1%	1.9%
Clustering Coeff. Watts-Strogatz (Av.)	0.016	0.016	0.016	0.016	3.2%	-1.9%	1.9%
Component Count Strong	3668	3702	3630	3673	0.9%	-1.0%	0.1%
Component Count Weak	202	195	207	201	-3.5%	2.5%	-0.5%
Component Mem-bers-Weak (Av.)	12.7	11.8	13.1	12.7	-6.8%	3.2%	0.7%
Average Distance	6.182	7.053	6.243	6.973	14.1%	1.0%	12.8%
Boundary Spanner (Av.)	0.223	0.229	0.224	0.230	3.0%	0.4%	3.5%
Interlockers (Av.)	0.018	0.015	0.017	0.019	-16.6%	-1.7%	6.3%
Connectedness	0.780	0.785	0.773	0.776	0.6%	-0.9%	-0.5%
Hierarchy	0.959	0.960	0.959	0.963	0.2%	0.0%	0.4%
Efficiency-Global	0.180	0.177	0.178	0.173	-1.8%	-1.3%	-3.8%
Upper Boundedness	0.400	0.434	0.396	0.412	8.7%	-0.9%	3.2%
Link Count Reciprocal	0.006	0.004	0.006	0.004	-31.0%	0.0%	-32.8%
Transitivity	0.027	0.027	0.027	0.027	1.5%	-0.7%	-1.1%
Triad Count (Av.)	0.255	0.308	0.247	0.249	20.6%	-3.3%	-2.5%
Clique Count (Av.)	0.578	0.535	0.558	0.493	-7.5%	-3.4%	-14.7%

This finding contrasts with the corpus statistics that we computed on the entity and relation level, where AR and CR induced the same trend with just different magnitudes. We assume that this divergence in findings is due to the fact that SNA is particularly sensitive to the connectivity and weight of individual nodes. These two node characteristics impact a node’s prominence and importance in the graph and thereby also the overall network structure. On the corpus level, nodes were only embedded in dyads (regular links), whereas in the resulting social network, if a node has dyadic links to more than one other unique node, this node’s degree (number of direct links) increases accordingly. While in corpus statistics, the impact of heavy “outliers” (hubs) can be diluted by computing averages – as we did in Tables II and IV - SNA handles such data points more appropriately. Furthermore contrary to our findings from corpus statistics is the fact that AR on average causes a greater change in the value of measures than CR does. We know, however, that CR is more effective in aggregating the data, which can be confirmed for link consolidation on the relation level as shown in row two of Table V. One explanation for this discrepancy is the fact that AR leads to the split-up of highly central yet unspecified nodes, such as collective nodes for “he” and “she”, into multiple nodes labeled as named entities and with considerably lower weights. This procedure seems to impact the network structure and respective measures more strongly than the merging of links onto unique links with higher weights does.

In a final step we looked at the identity (name) of the nodes that repeatedly ranked highest (top three nodes) among the measures presented in Table V. Table VI shows the top five results. The rankings for AR and AR plus CR are fairly similar to each other, and so are the ones for raw data and CR. The two top scoring nodes in the raw data and data after CR are pronouns, which are unlikely to present the actual agents who drive the dynamics of a system. The lists of top scoring nodes for AR and AR plus CR seem more meaningful for practical applications. Overall, the SNA results suggest that AR has a strong and desirable impact on highly important nodes, while CR supports the normalization of a wide range of less crucial links, which is less obvious from the SNA results but still crucial for increasing the data quality.

Table VI
Top scoring nodes per routine

rank	raw data	AR	CR	AR and CR
1	his	president	he	president
2	he	officials	his	officials
3	officials	U.S.	officials	U.S.
4	U.S.	Sonny Bono	U.S.	some
5	Sonny Bono	Heavenly Ski Resort	group	boss

V. Limitations and Conclusions

Our findings quantify the maximum impact that reference resolution, a widely used text preprocessing techniques, can have on entity and relation extraction as well as on social network analysis. The insights gained herein strongly depend on the underlying data. While the ACE data set is fairly large and was annotated by trained human coders, it still represents one specific type of texts data, namely news coverage. Correct spelling and grammar and a coherent writing style are typical characteristic of these data. More colloquial data, such as blogs and emails, might show a more diverse and error prone writing style, which can lead to different results. This assumption would need to be verified in subsequent studies. The findings on the joint impact of AR and CR are furthermore limited by the order of the application of both routines. We applied AR prior to CR. This strategy increases the occurrence of named entities throughout the text, which again increases the amount of information that CR can exploit when reasoning about possible identity matches of entity pairs. However, performing CR first might leave AR with a less confusing mass of entities to choose from. We did not employ the second strategy and therefore cannot make a judgment about the preferable order of applying AR and CR.

In our data, applying AR and CR together leads to the normalization, deduplication and personalization of about 75% of all entities. CR seems to contribute more strongly to this effect than AR does, while AR is a precondition for effective CR. On the relation level, up to 13% of the links are modified and up to 8% are reduced due to AR plus CR. These results are more moderate because the relation extraction was performed on the sentence level only. Running social network analysis on the data resulting from relation extraction leads to findings contrary to what was observed during corpus statistics analysis: AR and CR cause different directions in the change of network analytical measures, AR alters network analytical measures more strongly than CR does, and either technique leads to different nodes scoring highest across a plethora of standard SNA measures. Bringing the results from corpus statistics and social network analysis together, we conclude that CR is more effective on the quantitative data deduplication level, while AR is a more powerful technique for splitting up generic nodes into named nodes with adjusted weights. This implies that for social network analysis, AR is highly recommendable, but CR should also be applied in order to bring the network structure closer to the true underlying social structure. Taking this argument one step further, network analysts could use reference resolution to test if nodes that dominate the network structure and respective measures might in fact represent different social entities and therefore should be split up accordingly. In summary, using highly accurate reference resolution techniques can lead to significant and desirable changes in the data. AR and CR help to identify unique named entities and aggregating the empiric evidence per entity. The alterations that can be attributed to reference resolution are qualitatively and quantitatively meaningful: the identities of entities, relations, nodes, and edges change along with their statistical properties.

Acknowledgment

The authors are grateful to Jesse St. Charles from Carnegie Mellon University for his comments on this paper.

References

[1] J. Diesner and K. M. Carley, "Looking under the hood of machine learning algorithms for parts of speech tagging," Carnegie Mellon University, School of Computer Science, Institute for Software Research CMU-ISR-08-131R, 2008.

[2] C. Sidner, "Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse." vol. PhD Boston, MA: MIT, 1979.

[3] J. Hobbs, "Coherence and coreference," *Cognitive science,* vol. 3, pp. 67-90, 1979.

[4] J. Sowa, "Semantic Networks," in *Encyclopedia of Artificial Intelligence*, 2nd ed, S. C. Shapiro, Ed. New York, NY, USA: Wiley and Sons, 1992, pp. 1493 - 1511.

[5] M. Doerfel, "What Constitutes Semantic Network Analysis? A Comparison of Research and Methodologies," *Connections,* vol. 21, pp. 16-26, 1998.

[6] W. Woods, "What's in a link: Foundations for semantic networks," in *Representation and Understanding: Studies in Cognitive Science*, D. Bobrow and A. Collins, Eds. New York, NY: Academic Press, 1975, pp. 35-82.

[7] J. Diesner and K. M. Carley, "Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis," in *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and*

Illustrations, V. K. Narayanan and D. J. Armstrong, Eds. Harrisburg, PA: Idea Group Publishing, 2005, pp. 81-108.

[8] V. Krebs, "Mapping networks of terrorist cells," *Connections,* vol. 24, pp. 43-52, 2002.

[9] W. Van Atteveldt, *Semantic network analysis: Techniques for extracting, representing, and querying media content*. Charleston, SC: BookSurge Publishers, 2008.

[10] P. J. Carrington, J. Scott, and S. Wasserman, *Models and Methods in Social Network Analysis*: Cambridge University Press, 2005.

[11] A. McCallum, "Information extraction: distilling structured data from unstructured text," *ACM Queue,* vol. 3, pp. 48-57, 2005.

[12] J. Diesner and K. M. Carley, "Conditional Random Fields for Entity Extraction and Ontological Text Coding," *Journal of Computational and Mathematical Organization Theory,* vol. 14, pp. 248 - 262, 2008.

[13] M. Barthelemy and E. Chow, "Knowledge Representation. Issues in Semantic Graphs for Relationship Detection," in *AAAI Spring Symposium on AI Technologies for Homeland Security* Stanford, CA: AAAI Press, 2005, pp. 91-98.

[14] R. Bunescu and R. Mooney, "Statistical Relational Learning for Natural Language Information Extraction," in *Statistical Relational Learning*, L. Getoor and B. Taskar, Eds.: MIT, 2007, pp. 535 - 552.

[15] A. Culotta, A. McCallum, and J. Betz, "Integrating probabilistic extraction models and data mining to discover relations and patterns in text," in *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* New York, NY: Association for Computational Linguistics, 2006, pp. 296 - 303.

[16] K. Deemter and R. Kibble, "On Coreferring: Coreference in MUC and Related Annotation Schemes," *Computational Linguistics,* vol. 26, pp. 629-637, 2000.

[17] P. Denis and J. Baldridge, "Joint Determination of Anaphoricity and Coreference Resolution using Integer Programming," in *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, 2007, pp. 236-243.

[18] A. Haghighi and D. Klein, "Unsupervised Coreference Resolution in a Nonparametric Bayesian Model," in *Annual Meeting of the Association for Computational Linguistics*, 2007.

[19] A. Culotta, M. Wick, and A. McCallum, "First-Order Probabilistic Models for Coreference Resolution," in *North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, 2007, pp. 81-88.

[20] V. Ng, "Shallow semantics for coreference resolution," in *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2007, pp. 1689–1694.

[21] LinguisticDataConsortium, "Automatic Content Extraction."

[22] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, "The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation," in *4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 2004, pp. 837–840.

[23] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*: Prentice Hall, 2000.

[24] D. Roth and W. Yih, "Global Inference for Entities and Relations Identification via a Linear Programming Formulation," in *Statistical Relational Learning*, L. Getoor and B. Taskar, Eds. Boston, MA: MIT Press, 2007, pp. 535-552.

[25] P. A. Schrodt, Ö. Yilmaz, D. J. Gerner, and D. Hermick, "Coding Sub-State Actors using the CAMEO (Conflict and Mediation Event Observations) Actor Coding Framework," in *Annual Meeting of the International Studies Association* San Francisco, CA, 2008.

[26] J. A. Danowski, "Network Analysis of Message Content," *Progress in Communication Sciences,* vol. 12, pp. 198-221, 1993.

[27] K. M. Carley, D. Columbus, M. DeReno, J. Reminga, and I.-C. Moon, "ORA User's Guide 2008," Carnegie Mellon University, School of Computer Science, Institute for Software Research CMU-ISR-08-125, 2008.

[28] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*: Cambridge University Press, 1994.