

# Impact of Relation Extraction Methods from Text Data on Network Data and Analysis Results

Jana Diesner

University of Illinois at Urbana Champaign  
501 E Daniel Avenue  
61820 Champaign, IL  
[jdiesner@illinois.edu](mailto:jdiesner@illinois.edu)

Kathleen M. Carley

Carnegie Mellon University  
5000 Forbes Avenue  
15213 Pittsburgh, PA

[kathleen.carley@cs.cmu.edu](mailto:kathleen.carley@cs.cmu.edu)

## ABSTRACT

Many methods are available for extracting relational data from natural language text data. While these methods have been applied across corpora and domains, there is a lack of understanding of the differences in network structure and properties that result from employing different methods. We report on the comparison of relational data constructed by applying four commonly used methods – ranging from heavily manual to fully automated ones - to three large-scale, over-time corpora from different domains. Our comparison of the resulting data shows that there is little overlap between the networks per method. Ground truth data are partially resembled by analyzing the content of text bodies, but not at all by relying on meta-data only. We characterize the different perspectives of a network that are captured by each method, and suggest a strategy for combining these methods in order to gain a more holistic view on a network.

## Categories and Subject Descriptors

I.2.7 [Natural Language Processing]: *Text analysis*.

## General Terms

Measurement, Reliability.

## Keywords

Relation extraction, Evaluation, Method comparison.

## 1. INTRODUCTION

When network data are needed and text data are available as a source of information, network data can be extracted from text data. In computer science, this task is referred to as Relation Extraction (REX). Methods for going from texts to networks have been developed in different fields, mainly Artificial Intelligence (AI) [23], Natural Language Processing (NLP) and Computational Linguistics [19], social science [5; 15] and political science [14]. While these methods differ in their terminology, underlying theories and assumptions, degree of automation, evaluation methods and typical application areas, they overlap in that they exploit (a mixture of) lexical, syntactic, semantic, pragmatic, logical and statistical information (for a review see [10]), meta-data on the texts, and external information. Currently, the most accurate, efficient and scalable REX methods combine techniques from NLP, statistics and machine learning [18; 24]. One major decision involved in the end-to-end process of going from unstructured, natural language text data to network data is the selection of a REX method. Here, accuracy rates and time requirements are often used as decision criteria. However, even though accuracy and reliability rates are available for many of these methods, there is a lack of research on how these methods compare with respect to the structure and properties of the network data they generate. In this paper, we address this gap by providing an answer to the

following research question: How do the network data and network analysis results obtained by using different relation extraction methods compare to each other? In comparing four different relation extraction methods herein, we are not designing or hoping for convergence of the network structures and properties. Instead, our goal is to understand the differences and commonalities between the resulting data. This knowledge can contribute to our understanding of the different views on a network that relation extraction methods provide.

Why does this research matter? Having a better understanding of how relation extraction methods compare with respect to the data they generate contributes to the generalizability and transparency of these methods, increases the control that users and developers have over these analysis processes, and supports the drawing of reasonable and valid conclusions from findings.

## 2. METHODS

Four methods for constructing network data from text corpora are considered:

First, thesaurus-based text coding: the key component needed for this process is a thesaurus, which maps text terms to nodes, and sometimes also to node classes. Creating and adapting thesauri requires substantial human effort in terms of training and time. This process often involves a combination of NLP techniques such as computing (weighted) term frequencies to identify salient terms, adapting existing thesauri to a new corpus, content domain, genre and time frame, and using external knowledge sources such as word lists provided by subject matter experts. The resulting thesauri associate text terms with a node label and a node class. We refer to these thesauri as master thesauri herein because they serve as points of comparison for automatically generated thesaurus. The process of finding node labels involves reference resolution, i.e. associating different spellings of a concept with the same unique identifier. The node classes considered herein are: agent, organization, task, event, time, location, resources and knowledge. For some of these classes, we also distinguished specific from generic instances. The thesaurus creation for the datasets considered herein (described in the next section) took between two days and six weeks. Once the nodes have been identified there are several approaches for connecting them into links. Common methods for this purpose rely on (a combination of) proximal, syntactic, logical, and statistical information from text data. In this study, we use windowing; a widely used approach that connects any instances of nodes within a user-defined semantic unit and number words with each other [7]. Based on empirical tests, we chose a window size of seven within sentences [9]. In the following, this method is abbreviated as THES.

The second method resembles the first one, but uses a thesaurus constructed by applying probabilistic prediction models that we built via supervised machine learning based on Conditional Random Fields [9]. The accuracy of the employed models (predicting 45 states or classes) is 87.7% [9]. Again, one thesaurus was built per corpus, but the thesauri were constructed automatically and refined manually. This reduced the time requirements to one hour per one thousand articles for applying the prediction models plus time for manual cleaning and refinement. In the following, this method is abbreviated as AUTO-THEs. This study brings prediction models for entity extraction that were trained on certain annotated data, in this case the BBN corpus [25], into different application contexts for which no ground truth data is available. This is highly relevant as it resembles common, real-world analysis scenarios. Furthermore, with this study, we provide a comparison of manually and automatically constructed thesauri.

Third, network data were built from structured meta-data. This process disregards the content of text documents, but instead uses key words and index terms that humans and/ or algorithms have assigned to documents. For newswire data, for example, such meta-data are the key words per article. We operationalized link formation between meta-data entities as follows: two entities are linked if they co-occur as meta-data for the same document. This operationalization resembles the notion of windowing such that the network data constructed with the previous two text coding methods and those built from meta-data are based on the same notion of link formation. The advantage with network construction from meta-data is speed: once the meta-data are downloaded and organized in some structured form, such as a table or database, generating networks this way is basically a data retrieval task, which takes a couple of minutes. The limitation with this approach is that the assignment of meta-data entries to documents might not be transparent or documented. In the following, this method is abbreviated as META.

Fourth, we collaborated with subject matter experts (SME) on the Sudan data (described below) to build a tribal affiliation network for each calendar year from 2003 to 2008. This was an iterative process that involves the following steps: First, constructing initial network visualizations by using the THEs method; extracting organizations of subtype specific only and matching them against a list of tribes in the Sudan provided by the SMEs. Second, changing the network data according to the evaluation and feedback from the SMEs. We repeated this process until the SMEs evaluated the maps as representing the actual situation in the Sudan. The advantage with this approach is that it results in validated ground truth data, and this is the only ground-truth data set available for this study. In fact, for practical purposes, there are often no ground truth data on socio-technical networks available. This approach also involves two disadvantages: first, this process is expensive in terms of time and human resources: going through this process took several weeks. This amount of time is comparable to what is needed for constructing or cleaning thesauri by using the THEs method. Second, this process does not scale up, and is therefore only appropriate for generating datasets of small to moderate size. In the following, this method is abbreviated as SME.

The comparison of network data and analysis result in this paper is operationalized as follows: networks are compared with respect to the key entities that are identified according to selected network metrics. We do not provide comparisons on just the network metrics level since our prior work has shown

that any inaccuracies in reference resolution on node level can cause large error propagation rates on the link and network metrics level. In addition to these strategies for network comparison, the similarity of any pair of network data constructed with different methods is assessed by creating the intersection of these networks in terms of nodes and edges.

### 3. DATA

The selected relation extraction methods are applied to three large-scale, over time, open source corpora from different domains.

#### 3.1 Sudan

The first author put together the Sudan corpus by downloading relevant documents<sup>1</sup> from the LexisNexis Academic database, parsing, deduplicating and cleaning these documents based on data-driven rules and heuristics that she identified, evaluated and implemented, and managing the data in a relational database [9]. In total, the employed cleaning and deduplicating techniques reduced the corpus by 33.8% to 79,388 files. The text bodies were downloaded and stored along with meta-data. One type of meta-data are index terms, for which the categories, e.g. “country” or “city”, and values per category, e.g. “Sudan” or “Khartoum”, are defined and assigned by LexisNexis Academic without further documentation on this process. Other types of meta-data include the source and publication date of articles.

#### 3.2 Funding

The first author built the Funding corpus based on data from CORDIS, which provides information on the research proposals that have been funded by the European Union through the “Framework Programmes for Research and Technological Development”, short Framework Programmes (FPs) [6]. For this corpus, which is also managed in a relational database, several cleaning procedures were employed [for details see 9].

Per project, CORDIS specifies the name, affiliation, and contact information for the project coordinator (PC); a role equivalent to principal investigators. The same information is provided for each collaborator if applicable. One major challenge was co-reference resolution, i.e. consolidating the variations of references to people into a consistent and unique name per person. The first author developed a set of data-driven rules and heuristics, which she iteratively applied and evaluated. In total, 65.2% of the people entries were unique ( $N = 143,700$ ).

The unstructured, natural language text data per project can comprise a title, description and additional information. Meta-data include a project’s start and end date, costs, amount of funding awarded, completion status, and various key words and index terms, which can be assigned by CORDIS and/ or the authors.

The completeness of project entries in CORDIS varies per FP; with later FPs being more complete. Since incomplete network data can lead to strongly biased results [3], only FPs 4 to 6 are considered herein. For these FPs, the ratio of projects with at least one person specified exceeds 80% [6].

#### 3.3 Enron

As part of the investigations into Enron, FERC collected a total of 619,449 emails from 158 Enron employees, mainly senior managers, and publicly released these data to allow everybody

---

<sup>1</sup> In LexisNexis, “Sudan” was used as the search term, “major world publications” as data source, and “country” as category.

to understand this investigation. This dataset contains information about many individuals who were not involved in any of the actions that were subject of the Enron case.

The original version of the dataset had a variety of integrity problems. We started off building the CASOS Enron database by using the respective relational database from ISI [11]. The ISI researchers had cleaned the dataset by dropping blank, duplicated and junk emails, and emails that had been returned by the system due to transmission errors. This version had 252,759 emails from 151 distinct people. This dataset also required co-reference resolution: by default, the entities represent email addresses, not people. This leads to redundant nodes for people who use more than one email address. We have corrected for this issue by mapping e-mail addresses to individuals based on information from public sources [11]. We were able to map 1,234 email addresses to 557 distinct individuals for who we also know their actual name. In these refined data, the number of email addresses per person ranges from 1 to 17 with an average number of addresses per person of 2.2. The number of emails for which both a sender and at least one receiver can be mapped to a unique and disambiguated individual is 52,866.

Each email contains three types of information: First, explicit relational data from the email headers, i.e. the email addresses of the senders and receiver(s). We equally consider entries in the *to*, *cc*, and *bcc* fields as receivers. Second, text bodies, and third, meta-data such as time stamps and folder names.

### 3.4 Time slicing of datasets

For this project, time slicing was done based on calendar years (Sudan) and funding periods (Funding). For Enron, we decided to construct time slices around critical periods in Enron's history: the Enron crisis started to emerge in August 2001, when Jeffrey Skilling suddenly resigned as CEO and Kenneth Lay took over this position again. The crisis took off in October 2001, when Enron began to publicly report its losses. The stock market reacted with a sharp drop in prices for Enron shares; leading to the company's insolvency. Based on this timeline, we constructed three time periods for this study:

- May - June 2001 (6,091 emails): control case
- August – Sept. 2001 (3,711 emails): emergence of crisis
- October – December 2001 (11,042 emails): downfall

Taken together, the emails in these three time periods account for 41.0% of all emails in the CASOS Enron dataset.

### 3.5 Comparison of datasets

All three corpora feature text bodies and meta-data. Moreover, they all allow for constructing social networks and semantic networks. **Error! Reference source not found.** specifies the latter point and compares the versions of the datasets used herein along various dimensions.

Table 1: Summary and comparison of datasets

Dimension	Sudan Corpus	Funding Corpus	Enron Corpus
Genre/ domain	Newswire/ geo-political	Scientific writing	Emails/ business
Size	79,388 articles	43,276 proposals	20,844 emails
Time span	8 years	12 years	7 months
Social	Explicit: index terms	Explicit:	Explicit: emails

network	Implicit: text bodies	data on PIs	headers
Semantic network	Implicit: text bodies	Implicit: project descriptions	Implicit: email bodies

## 4. RESULTS

The prediction models trained on ground truth data were applied to the text bodies from all three datasets such that one thesaurus for entity extraction was built per corpus (AUTO-THESS). The first author evaluated the performance of the thesauri in the application scenarios; coming to the following conclusions across datasets:

1. For the majority of the entity classes supported by the prediction models ( $N = 44$  at most), instances are predicted with an accuracy that is high enough for being employable in practical applications to new datasets and domains.
2. No meaningful differences in prediction accuracy were observed for different publication times, genres and writing styles.
3. The auto-generated thesauri generalize better to new datasets and domains than master thesauri (THES method).
4. Creating and refining auto-generated thesauri is more efficient (in terms of time costs) and effective (in terms of entity coverage rate) than creating and refining master thesauri.
5. The prediction accuracy of classes seems to be independent of the number of instances per class in the application domain.
6. The auto-generated thesauri feature limitations with respect to prediction accuracy. Therefore, we recommend verifying and if needed correcting auto-generated thesauri.
7. Classes that perform low during formal model assessment (k-fold cross validation) are more likely to show low performance during application as well. However, classes with high accuracy during formal model assessment can return poor results in the application and vice versa.
8. Specific entities are predicted with a lower accuracy than a) generic entities and b) entities without a specificity value. This might be due to data sparsity, i.e. a lower number of specific than generic agents contained in the text data.
9. Prediction accuracy drops with cumulative frequency of the predicted entity, i.e. the number of times that an entity is observed in a particular class and – if applicable – further sub-categories, such as specificity and subtype.
10. Two main types of errors were observed for the auto-generated thesauri across all three application scenarios: First, terms that typically occur in lower case get assigned to the wrong category (mainly specific agents and organizations) if they occur in capitalized form. This might be due to data sparsity, and mainly happens if these terms occur at the beginning of a sentence, or when all letters of a term are capitalized, e.g. for acronyms and “yelling” in emails. These cases can be removed from the thesauri by comparing the spelling and part of speech of any two entities, outputting the cases that differ in capitalization only, and making a decision about them by either manually vetting them, or relying on the frequency counts, which are included in the auto-generated thesauri. Second, terms with a low frequency (less than ten, especially one to five) often involve chains of multiple entities or of relevant entities in conjunction with highly frequent, domain specific terms. These can be removed from the thesauri by disregarding

- suggestions with low frequencies. Again, this decision should be based on screening the thesaurus and identifying a suitable cut-off value.
11. Entries in the agent generic and organization generic classes tend to overlap for the case of references to groups, such as “students” or “workers”. In the CASOS standard thesauri, such entries also occur in either thesaurus category. For practical applications, it seems justifiable and efficient to merge these two classes.
- The comparison of the network data generated with different methods on the structural level and with respect to key entities lead to the following conclusions:
1. Ground truth data constructed by subject matter experts are hardly resembled by any automated methods that analyze text bodies, and even less so by exploiting existing meta-data from text corpora (META). This means that trying to reconstruct social network data from the content of text body will lead to largely incomplete networks.
  2. Networks extracted from text bodies by using auto-generated thesauri (AUTO-THES) resemble networks generated with master thesauri (THES) more strongly in terms of nodes and edges than vice versa.
  3. THES networks resemble meta-data networks more closely than AUTO-THES networks. This is because in this study, master thesauri were enhanced with information from the same sources that were used for defining the nodes in meta-networks. At the same time, auto-generated thesauri and meta-data networks are built from disjoint pieces of information, namely text bodies and meta-data on the texts.
  4. Agreements in structure and key entities are mainly impacted by two factors: First, network size: the larger a network, the higher is the chance that it resembles parts of network data constructed with other methods. This finding is relevant as it has been shown that network metrics can correlate with network size [1; 12; 13; 17]. Consequently, observed differences in these metrics across networks constructed with different methods might be independent of differences in the underlying network, but rather be a consequence of the network construction methods; and in the case of this study especially the link formation methods. Second, overlap in thesaurus content: similarity in the entities considered in the thesauri or for network construction strongly impacts the agreement in structure and key players.
  5. Structural agreements are always considerably higher on the node level than on the edge level. However, this finding is heavily impacted by the link formation methods.
  6. Meta-data networks (META) are less likely than text-based networks (THES, AUTO-THES) to suffer from co-reference resolution issues. This is mainly because somebody or some algorithm has already solved this issue. In contrast to the meta-data networks, both types of text based networks tend to retrieve single first names as key entities, which can be difficult to map to unique people with a first and last name.
  7. For social networks (agents and organizations) constructed from news wire data, meta-data networks are more suited for providing an overview on major international key entities and their relations, while the text-based networks are more appropriate for gaining a localized view on geo-political entities, and also for retrieving information about their culture.
  8. Meta-data networks retrieve more specific entities (in a qualitative, not quantitative sense) than the text-based networks. For the case of knowledge networks, meta-data networks return more informative key entities than the text-based networks, while text-based networks identify many common place terms as key entities.
  9. Overall, it seems recommendable to combine meta-data networks with text-based networks to cover both, the common or highly salient terms in a domain with more specific, domain dependent information. For this purpose, it might suffice to combine the networks built with auto-generated thesauri (AUTO-THES) with the meta-data networks plus any information from subject matter experts if available for the following reasons:
    - a. The AUTO-THES networks resemble the THES networks better than vice versa.
    - b. The AUTO-THES networks lead to similar types of key entities than the THES networks.
    - c. The THES networks already partially overlap with the meta-networks.

## 5. CONCLUSIONS AND LIMITATIONS

The knowledge gained with this research is limited by the data sets and the methodological choices we made:

### 5.1 Data level

Even though the Sudan corpus was collected through LexisNexis from a variety of sources, most of the texts are from newspapers and news magazines that appear in English. The biases that are contained in these sources are carried over to the extracted network data. Especially the analysis of meta-data had shown that one of these biases is a focus on high-profile politicians from the Western world. The CORDIS database might be incomplete, i.e. some funded project might be missing. Also, the database is incomplete for the listed projects. Moreover, the CORDIS database does not list rejected proposals, and no public source might provide this information. The Enron data are also likely to be incomplete as only the email archives from 158 people were collected, and people might not have stored all of their emails in these archives.

### 5.2 Methods level

**AUTO-THES:** Even though automated text coding (D2M process) speeds up computer-assisted text coding, it involves various weaknesses: entity extraction tools are more likely than humans to retrieve duplicates and near duplicates [2]. This was also observed in the application contexts. On the other hand, machine coding offers perfect intercoder-reliability (at least for non-probabilistic methods) and excludes accuracy losses due to fatigue and coding biases due to individual contextualization or interpretation of the data [21]. Yet another approach to achieve higher accuracy of the auto-generated thesauri without revising the thesauri for every new project would be to use more profound domain adaptation techniques [8; 16]. These techniques do not necessarily require the retraining of the prediction models, which is a time-costly process, but use statistical techniques to adjust a trained model to a new domain.

**Co-reference resolution:** This data cleaning technique was to validate and refine the master thesauri and auto-generated thesauri, to refine the network data, and to clean the datasets. We developed co-reference resolution rules in a data-driven fashion depending on the data. Alternatively, conducting reference resolution on the input text data prior to generating thesauri would solve this issue in the same way as it is solved

for meta-data networks, such that reference resolution is not pushed off to the thesaurus or network data level.

**Link types:** All approaches for extracting network data from texts used in this study treat links as untyped network constituents. Another valuable extension to this work would be the classification of links. In political science, the categorization of links is a state of the art process in event data coding [2; 22]. Machine-learned based methods for learning prediction models for link types have also been provided [4; 20].

**Link formation:** The findings are limited by the link formation approach, namely windowing, used for the extraction of relational data from text data. Our prior work has shown that windowing involves the risk of false positive links [9].

## 6. ACKNOWLEDGMENTS

This work was supported by the Army Research Institute (ARI) W91WAW07C0063, the Army Research Laboratory (ARL/CTA) DAAD19-01- 2-0009, the Air Force Office of Scientific Research (AFOSR) MURI FA9550-05-1-0388, and the Office of Naval Research (ONR) MURI N00014-08-11186. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ARI, ARL, AFOSR, ONR, or the United States Government.

## 7. REFERENCES

- [1] ANDERSON, B.S., BUTTS, C., and CARLEY, K.M., 1999. The interaction of size and density with graph-level indices. *Social Networks* 21, 3, 239-268.
- [2] BOND, D., BOND, J., OH, C., JENKINS, J., and TAYLOR, C., 2003. Integrated data for events analysis (IDEA): an event typology for automated events data development. *Journal of Peace Research* 40, 6, 733-745.
- [3] BORGATTI, S.P., CARLEY, K.M., and KRACKHARDT, D., 2006. On the robustness of centrality measures under conditions of imperfect data. *Social Networks* 28, 2, 124-136.
- [4] BUNESCU, R. and MOONEY, R., 2007. Statistical Relational Learning for Natural Language Information Extraction. In *Statistical Relational Learning*, L. GETTOOR and B. TASKAR Eds. MIT, 535 - 552.
- [5] CARLEY, K.M., 1993. Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological methodology* 23, 75-126.
- [6] CORDIS, Community Research and Development Information Service
- [7] DANOWSKI, J.A., 1993. Network Analysis of Message Content. *Progress in Communication Sciences* 12, 198-221.
- [8] DAUMÉ, H., 2007. Frustratingly easy domain adaptation. In *Conference of the Association for Computational Linguistics (ACL)*, 256.
- [9] DIESNER, J., forthcoming. Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts. Carnegie Mellon University.
- [10] DIESNER, J. and CARLEY, K., 2010. Relation Extraction from Texts (in German, title: Extraktion relationaler Daten aus Texten). In *Handbook Network Research (Handbuch Netzwerkforschung)* C. STEGBAUER and R. HÄUBLING Eds. Vs Verlag, 507-521.
- [11] DIESNER, J., FRANTZ, T.L., and CARLEY, K.M., 2005. Communication Networks from the Enron Email Corpus. "It's Always About the People. Enron is no Different". *Computational & Mathematical Organization Theory* 11, 3, 201-228.
- [12] FAUST, K., 2006. Comparing social networks: size, density, and local structure. *MetodoloÅ¡ki zvezki* 3, 2, 185-216.
- [13] FRIEDKIN, N.E., 1981. The development of structure in random networks: an analysis of the effects of increasing network density on five measures of structure. *Social Networks* 3, 1, 41-52.
- [14] GERNER, D., SCHRODT, P., FRANCISCO, R., and WEDDLE, J., 1994. Machine Coding of Event Data Using Regional and International Sources. *International Studies Quarterly* 38, 1, 91-119.
- [15] GLASER, B. and STRAUSS, A., 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, New York, NY.
- [16] GUPTA, R. and SARAWAGI, S., 2009. Domain adaptation of information extraction models. *ACM SIGMOD Record* 37, 4, 35-40.
- [17] MARSDEN, P.V., 1990. Network data and measurement. *Annual Review of Sociology*, 435-463.
- [18] MCCALLUM, A., 2005. Information extraction: distilling structured data from unstructured text. *ACM Queue* 3, 9, 48-57.
- [19] MIHALCEA, R.F. and RADEV, D.R., 2011. *Graph-based Natural Language Processing and Information Retrieval*. Cambridge Univ Press.
- [20] ROTH, D. and YIH, W., 2002. Probabilistic reasoning for entity and relation recognition. In *International Conf. on Computational Linguistics (COLING)*.
- [21] SCHRODT, P., 2001. Automated coding of international event data using sparse parsing techniques. In *International Studies Association*, Chicago, IL.
- [22] SCHRODT, P.A., YILMAZ, Ö., GERNER, D.J., and HERMICK, D., 2008. Coding Sub-State Actors using the CAMEO (Conflict and Mediation Event Observations) Actor Coding Framework. In *Proceedings of the Annual Meeting of the International Studies Association* (San Francisco, CA, March 2008 2008).
- [23] SOWA, J., 1992. Semantic Networks. In *Encyclopedia of Artificial Intelligence*, S.C. SHAPIRO Ed. Wiley and Sons, New York, NY, USA, 1493 - 1511.
- [24] VAN ATTEVELD, W., 2008. *Semantic network analysis: Techniques for extracting, representing, and querying media content*. BookSurge Publishers, Charleston, SC.
- [25] WEISCHEDEL, R. and BRUNSTEIN, A., 2005. BBN Pronoun Coreference and Entity Type Corpus. LDC2005T33. Linguistic Data Consortium, Philadelphia.