frontiers
in Digital Humanities

ORIGINAL RESEARCH
published: 16 November 2017
doi: 10.3389/fdigh.2017.00022

Using Semantic Linking to Understand Persons' Networks Extracted from Text

Alessio Palmero Aprosio[1], Sara Tonelli[1], Stefano Menini[1,2] and Giovanni Moretti[1]*

[1] Digital Humanities Research Unit, Center for Information and Communication Technology, Fondazione Bruno Kessler, Trento, Italy, [2] Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

In this work, we describe a methodology to interpret large persons' networks extracted from text by classifying cliques using the DBpedia ontology. The approach relies on a combination of NLP, Semantic web technologies, and network analysis. The classification methodology that first starts from single nodes and then generalizes to cliques is effective in terms of performance and is able to deal also with nodes that are not linked to Wikipedia. The gold standard manually developed for evaluation shows that groups of co-occurring entities share in most of the cases a category that can be automatically assigned. This holds for both languages considered in this study. The outcome of this work may be of interest to enhance the readability of large networks and to provide an additional semantic layer on top of cliques. This would greatly help humanities scholars when dealing with large amounts of textual data that need to be interpreted or categorized. Furthermore, it represents an unsupervised approach to automatically extend DBpedia starting from a corpus.

Keywords: persons' networks, semantic linking, DBpedia ontology, clique classification, natural language processing

OPEN ACCESS

Edited by:
Taha Yasseri,
University of Oxford, United Kingdom

Reviewed by:
Fabian Flöck,
Leibniz Institut für Sozialwissenschaften (GESIS), Germany
Juan Julián Merelo,
University of Granada, Spain
David Laniado,
Eurecat (Spain), Spain

****Correspondence:***
Alessio Palmero Aprosio
aprosio@fbk.eu

Specialty section:
This article was submitted to Big Data, a section of the journal Frontiers in Digital Humanities

Received: *28 July 2017*
Accepted: *23 October 2017*
Published: *16 November 2017*

Citation:
Palmero Aprosio A, Tonelli S, Menini S and Moretti G (2017) Using Semantic Linking to Understand Persons' Networks Extracted from Text. Front. Digit. Humanit. 4:22. doi: 10.3389/fdigh.2017.00022

1. INTRODUCTION

In recent years, humanities scholars have faced the challenge of introducing information technologies in their daily research activity to gain new insight from historical sources, literary collections, and other types of corpora, now available in digital format. However, to process large amounts of data and browse through the results in an intuitive way, new advanced tools are needed, specifically designed for researchers without a technical background. Especially scholars in the areas of social sciences or contemporary history need to interpret the content of an increasing flow of information (e.g., news, transcripts, and political debates) in short time, to quickly grasp the content of large amounts of data and then select the most interesting sources.

An effective way to highlight semantic connections emerging from documents, while summarizing their content, is a network. To analyze concepts and topics present in a corpus, several approaches have been successfully presented to model text corpora as networks, based on word co-occurrences, syntactic dependencies (Sudhahar et al., 2015), or Latent Dirichlet Allocation (Henderson and Eliassi-Rad, 2009). While these approaches focus mainly on concepts, other information could be effectively modeled in the form of networks, i.e., *persons*. Indeed, persons' networks are the focus of several important research projects, for instance, Mapping the Republic of Letters,[1]

[1] http://republicofletters.stanford.edu/.

where connections between nodes have been manually encoded as metadata. However, when scholars need to manage large amounts of textual data, new challenges related to the creation of persons' networks arise. Indeed, the process must be performed automatically, and since networks extracted from large amounts of data can include thousands of nodes and edges, the outcome may be difficult to read. While several software packages have been released to display and navigate networks, an overview of the content of large networks is difficult to achieve. Furthermore, this task also poses a series of technical challenges, for example, the need to find scalable solutions, and the fact that, although single components to extract persons' networks from unstructured text may be available, they have never been integrated before in a single pipeline nor evaluated for the task.

In this work, we present an approach to extract persons' networks from large amounts of text and to use Semantic Web technologies for classifying clusters of nodes. This classification relies on categories automatically leveraged from DBpedia, proving an effective interplay among Natural Language Processing, Semantic Web technologies, and network analysis. Through this process, interpretation of networks, the so-called *distant reading* (Moretti, 2013), is made easier. We also analyze the impact of persons' disambiguation and coreference resolution on the task. An evaluation is performed both on English and on Italian data, to assess whether there are differences depending on the language, on the domains covered by the two corpora, or on the different performance of NLP tools.

The article is structured as follows: in Section 2, we discuss past works related to our task, while in Section 3, we provide a description of the steps belonging to the proposed methodology. In Section 4, the experimental setup and the analyzed corpus are detailed, while in Section 5, an evaluation of node and clique classification is provided and discussed. In Section 6, we provide details on how to obtain the implemented system and the dataset, and finally we draw some conclusions and discuss future work in Section 7.

2. RELATED WORK

This work lies at the intersection of different disciplines. It takes advantage of studies on graphs, in particular research on the proprieties of cliques, i.e., groups of nodes with all possible ties among themselves. Cliques have been extensively studied in relation to social networks, where they usually represent social circles or communities (Grabowicz et al., 2013; Jin et al., 2013; Mcauley and Leskovec, 2014). Although we use them to model co-occurrence in texts and not social relations, the assumption underlying this work is the same: the nodes belonging to the same clique share some common properties or categories, which we aim at identifying automatically, using the Linked Open Data.

This work relies also on past research analyzing the impact of preprocessing, in particular coreference resolution and named entity disambiguation, on the extraction of networks from text. The work presented in Diesner and Carley (2009) shows that anaphora and coreference resolution have both an impact on deduplicating nodes and adjusting weights in networks extracted from news. The authors recommend to apply both preprocessing steps to bring the network structure closer to the underlying social structure. This recommendation has been integrated in our processing pipeline, when possible.

The impact of named entity disambiguation on networks extracted from e-mail interactions is analyzed in Diesner et al. (2015). The authors argue that disambiguation is a precondition for testing hypotheses, answering graph-theoretical and substantive questions about networks, and advancing network theories. We base our study on these premises, in which we introduce a mention normalization step that collapses different person mentions onto the same node if they refer to the same entity.

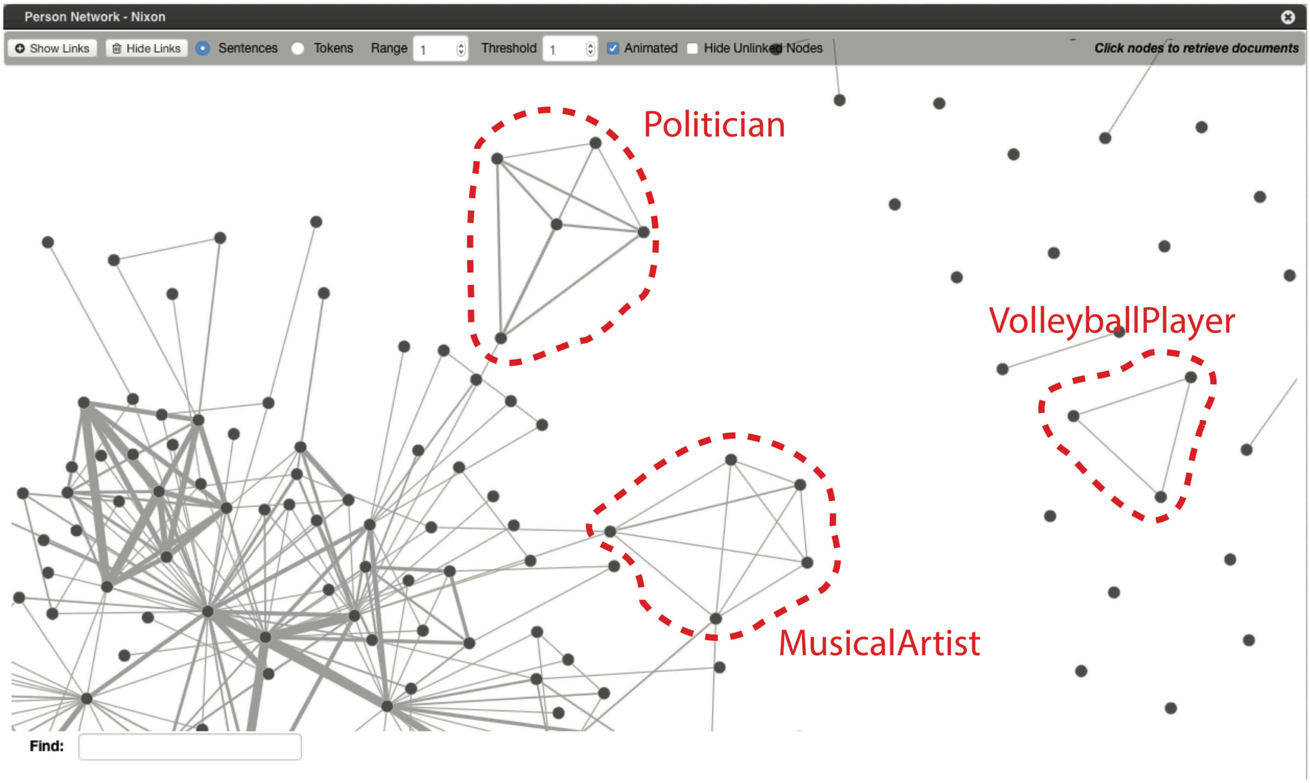
Kobilarov et al. (2009) describe how BBC integrates data and links documents across entertainment and news domains by using Linked Open Data. Similarly, in Özgür et al. (2008), Reuters News articles are connected in an entity graph at document-level: people are represented as vertices, and two persons are connected if they co-occur in the same article. The authors investigate the importance of a person using various ranking algorithms, such as PageRank. In Hasegawa et al. (2004), a similar graph of people is created, showing that relations between individuals can be guessed also connecting entities at sentence-level, with high precision and recall. In this work, we extract persons' networks in a similar way, but we classify groups of highly connected nodes rather than relations.

In Koper (2004), the Semantic Web is used to get a representation of educational entities, to build self-organized learning networks, and go beyond course and curriculum centric models. The *Trusty* algorithm (Kuter and Golbeck, 2009) combines network analysis and Semantic Web to compute social trust in a group of users using a particular service on the Web.

3. METHODOLOGY

We propose and evaluate a methodology that takes a corpus in plain text as input and outputs a network, where each *node* corresponds to a person and an *edge* is set between two nodes if the two persons are co-occurring inside the same sentence. Within the network, *cliques*, i.e., maximum number of nodes who have all possible ties present among themselves are automatically labeled with a category extracted from DBpedia. In our case, cliques correspond to persons who tend to occur together in text, for which we assume that they share some commonalities or the same events. The goal of this process is to provide a comprehensive overview of the persons mentioned in large amounts of documents and show dependencies, overlaps, outliers, and other features that would otherwise be hard to discern. A portion of a network with three highlighted cliques is shown in **Figure 1**.

The creation of a persons' network from text can be designed to model different types of relations. In case of novels, networks can capture dialog interactions and rely on the conversations between characters (Elson et al., 2010). In case of e-mail corpora (Diesner et al., 2015), edges correspond to emails exchanged between sender and addressee. Each type of interaction must be recognized with an *ad hoc* approach, for instance, using a tool that identifies direct speech in literary texts. On the contrary, our goal is to rely on a general-purpose methodology, therefore our approach to network creation is based on simple co-occurrence,



Person Network - Nixon

Show Links | Hide Links | ● Sentences ○ Tokens | Range 1 | Threshold 1 | ☑ Animated □ Hide Unlinked Nodes

Click nodes to retrieve documents

Find:

FIGURE 1 | Visualization of a large persons' network with examples of labeled cliques.

similar to existing approaches to the creation of concept networks (Veling and Van Der Weerd, 1999). In the following subsections, we detail the steps building our approach, displayed in **Figure 2**.

3.1. Preprocessing

Each corpus is first processed with a pipeline of NLP tools. The goal is to detect persons' names in the documents and link them to DBpedia. Since our approach supports both English and Italian, we adopt two different strategies, given that the NLP tools available for the two languages are very different and generally achieve better performance on English data. For English, we use the PIKES suite (Corcoglioniti et al., 2016): it first launches the Stanford Named Entity Recognizer (Finkel et al., 2005) to identify persons' mentions in the documents (e.g., "J. F. Kennedy," "Lady Gaga," etc.), and then the Stanford Deterministic Coreference Resolution System (Manning et al., 2014) to set coreferential chains within each document. For instance, the expressions "J. F. Kennedy," "J. F. K.," "John Kennedy," and "he" may all be connected because they all refer to the same person. For Italian, instead, no tool for coreference resolution is available, therefore only NER is performed, using the Tint NLP suite (Palmero Aprosio and Moretti, 2016).

Then, for both languages we run DBpedia Spotlight (Daiber et al., 2013) and the Wiki Machine (Palmero Aprosio and Giuliano, 2016) to link the entities in the text to the corresponding DBpedia pages.[2] In particular, we consider only links that overlap with the NER annotation and belong to the `Person` category. We combine the output of the two tools, since past works proved that this outperforms the performance of single linking systems (Rizzo and Troncy, 2012).

In case of mismatch between the output of the two linking annotations, the confidence values (between 0 and 1, provided by both systems) are compared, and only the more confident result is considered. At the end of preprocessing, we obtain for each document a list of (coreferring) persons' mentions linked to DBpedia pages.

3.2. Linking Filter

To improve linking precision, a filtering step based on Semantic Web resources has been introduced. It is applied to *highly ambiguous entities*, because it is very likely that they are linked to the wrong Wikipedia page, so it may be preferable to ignore them during the linking process. An entity should be ignored if the probability that it is linked to a Wikipedia page—calculated as described in Palmero Aprosio et al. (2013a)—is below a certain threshold. For instance, the word *Plato* can be linked to the philosopher, but

[2] DBpedia Spotlight has a reported accuracy of 0.85 on English and 0.78 on Italian. As for the Wiki Machine, the reference paper reports Precision 0.78, Recall 0.74, and F1 0.76 on English (no evaluation provided for Italian).

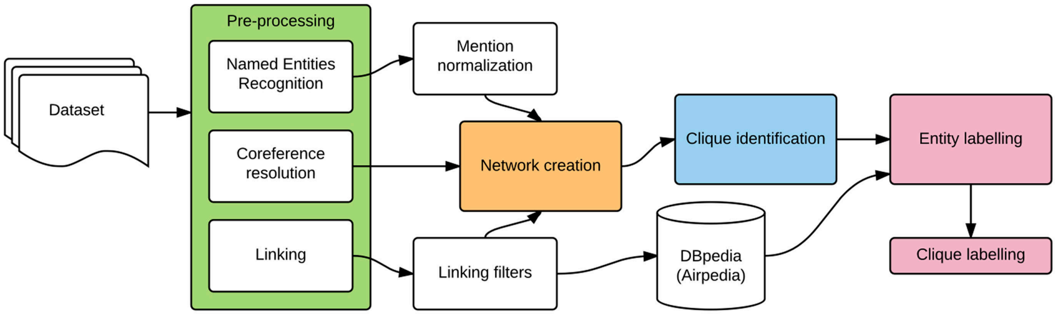


FIGURE 2 | Workflow of the whole system.

also to an actress, *Dana Plato*, a racing driver, *Jason Plato*, and a South African politician, *Dan Plato*. However, the probability that *Plato* is linked to the philosopher page is 0.93, i.e., the link to the philosopher is probably always right. That value is calculated considering—in Wikipedia—both the number of links referring to that entity, and the semantics of the context extracted from the text surrounding the linked entity. In some cases, thresholds are very low, especially for common combinations of name–surname. For example, *Dave Roberts* can be linked to 15 different Wikipedia pages, all of them having similar thresholds (0.19 for the outfielder, 0.14 for the pitcher, 0.06 for the broadcaster, 0.04 for the Californian politician mentioned in Kennedy's speeches, etc.). We manually checked some linking probabilities and set the threshold value to 0.2, so that if every possible page that can be linked to a mention has a probability <0.2, the entity is not linked. The impact of this step on the general task is reported in **Table 1**.

TABLE 1 | Number of nodes and cliques in the networks with and without mention normalization (MN) and coreference resolution (COREF—only for English).

	Dataset	w/o MN	MN
Number of nodes	NK	4,754	4.261
Number of nodes	Adige	28,644	19,133
Number of cliques			
w/o COREF	NK	720 (4.62)	683 (4.60)
COREF	NK	1,005 (4.91)	869 (4.80)
w/o COREF	Adige	14,762 (5.23)	6,294 (5.12)

In brackets, the average number of entities for each clique.

3.3. Network Creation

The goal of this step is to take in input the information extracted through preprocessing and filtering and produce a network representing person co-occurrences in the corpus. We assume that persons correspond to nodes and edges express co-occurrence, therefore we build a person–person matrix by setting an edge every time two persons are mentioned together in the same sentence.[3]

A known issue in network creation is name disambiguation, i.e., identifying whether a set of person mentions refers to one or more real-world persons. This task can be very difficult because it implies understanding whether spellings of seemingly similar names, such as "Smith, John" and "Smith, J.," represent the same person or not. The given problem can get more complicated, especially when people are named with diminutives (e.g., "Nick" instead of "Nicholas"), acronyms (e.g., "J.F.K.") or inconsistently spelled.

We tackle this problem with a **mention normalization** step based on a set of rules for English and Italian, dealing both with single- and multiple-token entities. Specifically, entities comprising more than one token (i.e., complex entities) are collapsed onto the same node if they show a certain amount of common tokens (e.g., "John F. Kennedy" and "John Kennedy"). The approach is similar to the *first initial* method that proved to reach 97% accuracy in past experiments (Milojević, 2013). As for simple entities (i.e., composed only of one token), they can be either proper names or surnames. To assess which simple entity belongs to which category, two lists of first and family names are extracted from biographies in Wikipedia, along with their frequency: a token is considered as a family name if it appears in the corresponding list and it does not appear in the first name list. Tokens not classified as surnames are ignored and not included in the network. Tokens classified as surnames, instead, are merged with the node corresponding to the most frequent complex entity containing such surname. The extraction of name and surname lists is performed using information included in infoboxes: in the English Wikipedia, the name and surname of a person are correctly split in `DEFAULTSORT`; in Italian, that information is included in `Persondata`.

For example, the single mentions of "Kennedy" are all collapsed onto the "John Fitzgerald Kennedy" node, if it is more frequent in the corpus than any other node containing the same surname such as "Robert F. Kennedy," "Ted Kennedy," etc. Normalization is particularly effective to deal with distant mentions of the same person in a document, because in such cases coreference tends to fail. It is also needed for documents in which ambiguous forms cannot be mapped to an extended version, for example, when only "Kennedy" is present. Finally, it is very effective on Italian, since for this language there is no coreference resolution tool. After mention normalization, the network has less nodes but it is more connected than the original version without normalization (see **Table 2**).

[3] Even if the sentence window is arbitrary, it is common to consider this boundary also when manually annotating relations in benchmarks (Mitchell et al., 2002; Hasegawa et al., 2004).

3.4. Clique Identification and Labeling

The last steps of the process include the identification of cliques, i.e., clusters of nodes with all possible ties among themselves (see **Figure 1**), and their classification by assigning a semantic category covering all nodes included in the clique. In case of small datasets, existing algorithms can quickly find all maximal cliques inside a network (a maximal clique is a clique that cannot be enlarged by adding a vertex). The most efficient one is the Bron–Kerbosch clique detection algorithm (Bron and Kerbosch, 1973). Unfortunately, the algorithm takes exponential time $O(3^{n/3})$ (being n the number of vertices in the network), which means that it quickly becomes intractable when the size of the network increases. Since in our scenario we are not interested in listing *every* maximal clique, but we can instead limit the size of the cliques to a fixed value k (that can be arbitrary big, for example, 10), the execution time drops to $O(n^k k^2)$, that is polynomial (Downey and Fellows, 1995).

Clique labeling is performed according to the following algorithm. Let C be the set of cliques to be labeled. For each clique $c \in C$, let c_i, $i = (1 \dots k_c)$ be the nodes belonging to c (note that we extract cliques of different sizes, thus we denote with k_c the size of the clique c). For each node c_i previously linked to a Wikipedia page (see Sections 3.1 and 5), we extract the corresponding DBpedia classes using Airpedia (Palmero Aprosio et al., 2013b). This system was chosen because it extends DBpedia coverage, classifying also pages that do not contain an infobox and exploiting cross-lingual links in Wikipedia. This results in a deeper and broader coverage of pages w.r.t. DBpedia classes. Let class (c_i) be the set of DBpedia classes associated with an entity $c_i \in c$. Note that class $(c_i) = \emptyset$ for some c_i, as only around 50% of the entities can be successfully linked (see last column of **Table 2**).

For each clique, we define the first frequency function F' that maps each possible DBpedia class to the number of occurrences of that class in that clique. For example, the annotated clique

$$\begin{gathered}\text{Gifford Pinchot} \rightarrow \texttt{Governor}\\ \text{Theodore Roosevelt} \rightarrow \texttt{President}\\ \text{Wendell Willkie} \rightarrow \left[\texttt{none}\right]\\ \text{Franklin Roosevelt} \rightarrow \texttt{President}\end{gathered}$$

will result in

$$\begin{gathered}F'(\texttt{Governor}) = 1\\ F'(\texttt{President}) = 2.\end{gathered}$$

As DBpedia classes are hierarchical, we compute the final frequency function F by adding to F' the ancestors for each class. In our example, as `Governor` and `President` are both children of `Politician`, F will result in

$$\begin{gathered}F(\texttt{Governor}) = 1\\ F(\texttt{President}) = 2\\ F(\texttt{Politician}) = 3.\end{gathered}$$

Since in our task we focus on persons, we only deal with the classes dominated by `Person` (we ignore the `Agent` class, along with `Person` itself). Finally, we pick the class that has the highest frequency and extend the annotation to the unknown entities. In the example, *Wendell Willkie* would be classified as `Politician`. The same class is also used to guess what the people in the clique have in common, i.e., a possible classification of the whole clique, to help the *distant reading* of the graph.

TABLE 2 | Evaluation of node and clique classification (HAE means "highly ambiguous entities").

Data	Experiment	P	R	F_1	# entities
NK	Baseline (Politician)	0.807	0.491	0.611	347/347
NK	Node classification	0.689	0.481	0.566	245/347
NK	Extension to non-linked	0.617	0.578	0.597	245/347
NK	Node classification, no HAE	0.870	0.460	0.602	176/347
NK	Extension to non-linked, no HAE	0.738	0.632	0.681	347/347
NK	Clique classification	0.677	0.768	0.720	
Adige	Baseline (context)	0.802	0.488	0.607	486/486
Adige	Node classification	0.891	0.256	0.398	154/486
Adige	Extension to non-linked	0.911	0.625	0.742	486/486
Adige	Clique classification	0.930	0.485	0.637	

4. EXPERIMENTAL SETUP

4.1. Evaluation Methodology

We evaluate our approach on two corpora:

- The corpus of political speeches uttered by Nixon and Kennedy (NK) during 1960 presidential campaign.[4] It contains around 1,650,000 tokens (830,000 by Nixon and 815,000 by Kennedy).
- A corpus extracted from articles published on the Italian newspaper L'Adige[5] between 2011 and 2014, containing 9,786,625 tokens. To increase the variability of the news content and have a balanced dataset, we retrieve the documents from different news sections (e.g., Sports, Politics, and Events).

The corpus is first pre-processed as described in Section 3.1. Then, the recognized entities are linked and mention normalization (MN) is performed. On the English data we also run coreference resolution (COREF) (see Section 3.1). We show in **Table 1** the impact of these two processes on the network dimension and on the number of extracted cliques.

Clique identification is performed by applying the Bron–Kerbosch clique detection algorithm (see Section 3.4), using the implementation available in the JGraphT package.[6] After this extraction, we only work on cliques having at least 4 nodes, as smaller cliques would be too trivial to classify. **Table 3** lists the number of cliques grouped by size.

Mention normalization reduces the number of nodes because it collapses different mentions onto the same node. Consequently, the number of cliques decreases (see **Table 1**).

[4] The transcription of the speeches is available online by John T. Woolley and Gerhard Peters, The American Presidency Project (http://www.presidency.ucsb.edu/1960_election.php).

[5] http://www.ladige.it/.

[6] http://jgrapht.org/.

TABLE 3 | Number of cliques grouped by size.

Dataset/size	4	5	6	7	8	9	10	11	12	14	15	16	17	19	20	23
NK	211	158	100	66	39	17	7	5	3	0	0	1	1	1	1	0
Adige	177	120	89	33	40	5	3	1	0	2	2	0	0	0	0	1

Coreference resolution, instead, does not have any impact on the network dimension, but it increases the number of edges connecting nodes, resulting in an increment of the number of cliques and also of their dimension. The evaluation presented in the remainder of this article on English data is based on a system configuration including both mention normalization and coreference resolution. For Italian, only mention normalization is performed.

4.2. Gold Standard Creation

Since the goal of this work is to present and evaluate a methodology to assign categories to cliques and make large persons' networks more readable, we first create a gold standard with two annotated layers, one at *node* and one at *clique* level. This data set includes 184 cliques randomly extracted from the clique list (see Section 3.4): 84 from the NK corpus and 100 from Adige.

First, each node in the clique is manually annotated with one or more classes from the DBpedia ontology (Lehmann et al., 2015) expressing the social role of the person under consideration. For example, *Henry Clay* is annotated both as `Senator` and `Congressman`. For many political roles, the ontology does not contain any class (for instance, *Secretary*). In that case, the person is labeled with the closest more generic class (e.g., *Politician*). Then, for each clique, we identify the most specific class (or classes) of the ontology including every member of the group. The shared class is used as label to define the category of the clique. For example, a clique can be annotated as follows:

$$\begin{gathered}\text{John Swainson} \rightarrow \texttt{Governor}\\ \text{G. Mennen Williams} \rightarrow \texttt{Governor}\\ \text{Thaddeus Machrowicz} \rightarrow \texttt{Congressman}\\ \text{Jim O'Hara} \rightarrow \texttt{Congressman}\\ \text{Pat McNamara} \rightarrow \texttt{Senator}\\ [\textit{whole clique}] \rightarrow \texttt{Politician}.\end{gathered}$$

In case no category covering all nodes exists, the `Person` class is assigned. For instance, a clique containing 3 nodes labeled as `Journalist` and 2 nodes as `President` is assigned the `Person` class.

The gold standard contains overall 833 persons (347 from NK, 486 from Adige) grouped into 184 cliques, only 27 of which are labeled with the *Person* category (13 from NK, 14 from Adige). This confirms our initial hypothesis that nodes sharing the same clique (i.e., persons who tend to be mentioned together in text) show a high degree of commonality. All entities in the gold standard are assigned at least one category. Since this task is performed by looking directly at the DBpedia ontology, also persons who are not present in Wikipedia are manually labeled. In case a node is ambiguous (e.g., six persons named *Pat McNamara* are listed in Wikipedia), the annotator looks at the textual context(s) in which the clique occurs to disambiguate the entity.

5. RESULTS

In **Table 2**, we report different stages of the evaluation performed by comparing the system output with the gold standards presented in the previous subsection. We also compare our performance with a competitive baseline: for NK, we assign to each clique the **Politician** category, given that this pertains to the domain of the corpus. For Adige, we select the most probable category by article affinity: `Athlete` for sport section, `Artist` for cultural articles, `Politician` for the remaining ones.

We first evaluate the classification of the single nodes ("*node classification*") by comparing the category assigned through linking with DBpedia Spotlight and the Wiki Machine to the class labels in the gold standard. Since our methodology assigns a category to a clique even if not all nodes are linked to a Wikipedia page, we evaluate also the effect of inheriting the clique class at node level (see row "*Extending to non-linked entities*").

Besides, we assess the impact of "*highly ambiguous entities*" on node classification, and the effect of removing them from the nodes to be linked ("*without highly ambiguous entities*"). For instance, we removed from the data the node of "Bob Johnson," which may refer to 21 different persons (see Section 3.2 for details). Note that we report the results only for English, since this step had no effect on the Italian data, containing no person mention with a relevance < 0.2. The last line for each dataset in **Table 2** shows the performance of the system on guessing the shared class for the entire clique.

For each entity that needs to be classified, the evaluation is performed as proposed by Melamed and Resnik (2000) for a similar hierarchical categorization task. **Figure 3** shows an example of the evaluation. The system tries to classify the entity *Dante Fascell* and maps it to the ontology class `Governor`, while the correct classification is `Congressman`. The missing class (question mark) counts as a false negative (*fn*), the wrong class (cross) counts as a false positive (*fp*), and the correct class (tick) counts as a true positive (*tp*). As in this task we classify only people, we do not consider the true positives associated to the `Person` and `Agent` classes.[7] In the example above, classification of *Dante Fascell* influences the global rates by adding 1 *tp*, 1 *fn*, and 1 *fp*. Once all rates are collected for each classification, we calculate standard precision (p), recall (r), and F_1.

Results in **Table 2** show some differences between the English and the Italian dataset. With NK, that deals with people who

[7] See http://mappings.dbpedia.org/server/ontology/classes/ for a hierarchical representation of the DBpedia ontology classes.

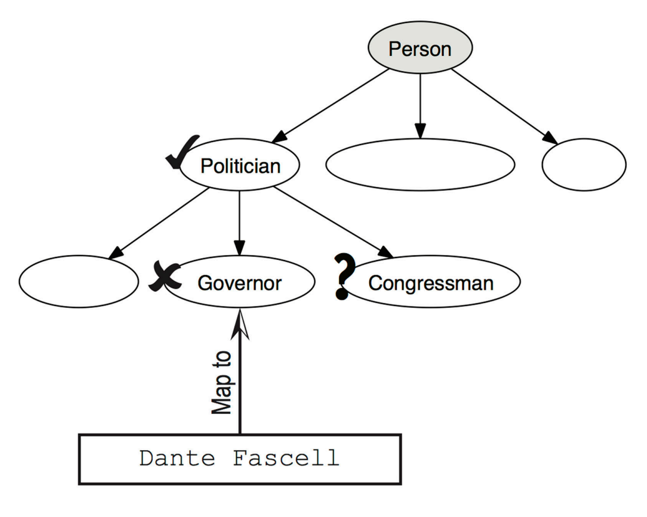


FIGURE 3 | Description of the evaluation.

lived in the sixties, the performance of node classification suffers from missing links, depending on the incomplete coverage of DBpedia Spotlight and the Wiki Machine, but also on the fact that some entities are not present in Wikipedia. However, this configuration achieves a good precision. In terms of F_1, extending the class assigned to the clique also to non-linked entities yields a performance improvement, due to better recall. Removing highly ambiguous entities is extremely beneficial because it boosts precision as expected, especially in combination with the strategy to extend the clique class to all underlying nodes. The setting based on this combination is the best performing one, achieving an improvement with respect to basic node classification both in precision and in recall. On this corpus, the baseline assigning the `Politician` label to all nodes is very competitive because of the domain. Based on the best performing setting for node classification, we evaluated the resulting clique classification, with the goal of assigning a category to clusters of interconnected nodes and easing the network comprehension. Results show that the task achieves good results and, even if not directly comparable, classification performance is higher than on single nodes.

In the Adige corpus, precision is higher than in NK: the persons mentioned in this dataset are in most of the cases still living, therefore they are present in Wikipedia more often than the persons mentioned in NK in 1960. On the contrary, recall is lower. We investigated this issue and discovered that entities in DBpedia are often not classified with the most specific class. For example, Mattia Pellegrin is a cross country skier and was annotated as `CrossCountrySkier` by our annotators. On the contrary, in DBpedia the entity `Mattia_Pellegrin` is classified as `Athlete`, thus this was the label assigned by our system. Following the evaluation described in **Figure 3**, our system is penalized as it misses both `WinterSportPlayer` and `CrossCountrySkier`. For this reason, in classifying Mattia Pellegrin, the system gets 1 *tp* and 2 *fn*.

Being able to assign classes to cliques, even if not all nodes are linked, our approach has a high potential in terms of coverage. Indeed, it can cover entities that are not in Wikipedia (and in DBpedia), by guessing their class using DBpedia categories. In general terms, it may be used also to automatically extend DBpedia with new person entities. Specifically, we classified 171 new entities in NK (p = 0.738 and F_1 = 0.681) and 332 entities in Adige (p = 0.911 and F_1 = 0.742), for a total of 503. Given that the gold standard includes 833 entities, this means that on average 60% of entities in the two datasets (503 out of 833) are not present in Wikipedia (or are too ambiguous, see description of "highly ambiguous entities" in Section 3.2), and our system is capable of assigning them a DBpedia category. Our gold standard is relatively small, but if this step is launched on a large amount of data, it has the potential to significantly extend DBpedia with unseen entities, for example, those living in the past who are not represented in the knowledge base. On the other hand, we are aware that the way Wikipedia is built and edited can affect the outcome of this work. In particular, Wikipedia Western and English bias must be taken into account when using this kind of approaches for studies in the digital humanities (e.g., cultural analytics), because certain persons' categories and nationalities are more present than others.

6. DATASETS AND TOOL

The tool performing the workflow described in this paper is written in Java and released on GitHub[8] under the GPL license, version 3. On our GitHub page one can find:

- the dataset containing the original Nixon and Kennedy speech transcriptions (released under the NARA public domain license) along with the linguistic annotations applied in the preprocessing step (in NAF format (Fokkens et al., 2014), see Section 3.1);
- the annotated cliques for both datasets (NK and Adige).
- Unfortunately, the Adige corpus is not publicly released, therefore we cannot make it available for download.

7. CONCLUSION AND FUTURE WORK

In this work, we presented an approach to extract persons' networks from large amounts of textual data based on co-occurrence relations. Then, we introduced a methodology to identify cliques and assign them a category based on DBpedia ontology. This additional information layer is meant to ease the interpretation of networks, especially when they are particularly large.

We discussed in detail several issues related to the task. First of all, dealing with textual data is challenging because persons' mentions can be variable or inconsistent, and the proposed approach must be robust enough to tackle this problem. We rely on a well-known tool for coreference resolution and we perform mention normalization, so that all mentions referring to the same entity are recognized and assigned to the same node. We also introduced a filtering strategy based on information retrieved from Semantic Web resources, to deal with highly ambiguous entities.

Finally, we presented and evaluated a strategy to assign a category to the nodes in a clique and then, by generalization, to

[8] https://github.com/dkmfbk/cliques.

the whole clique. The approach yields good results, especially in terms of precision, both at node and at clique level. Furthermore, it is able to classify entities that are not present in Wikipedia/DBpedia and could be also used to enrich other knowledge bases, for example, Wikidata, without any supervision. The data manually annotated for the gold standards confirm the initial hypothesis that co-occurrence networks based on persons' mentions can provide an interesting representation of the content of a document collection, and that cliques can effectively capture commonalities among co-occurring persons. To the best of our knowledge, this hypothesis was never proved before, and the clique classification task based on DBpedia ontology is an original contribution of this work.

In the future, we plan to integrate this methodology in the ALCIDE tool (Moretti et al., 2016), which displays large persons' networks extracted from text but suffers from a low readability of the results. We also plan to improve and extend nodes and cliques classification, for instance, by applying clique percolation (Palla et al., 2005), a method used in Social Media analysis to discover relations between communities (Gregori et al., 2011). Another research direction will deal with almost-cliques (Pei et al., 2005) or node clusters with high (but not maximal) connectivity, so as to increment the coverage of our approach by including more entities. Finally, we would like to exploit the links connecting different Wikipedia biographies to cross-check the information automatically acquired from cliques and investigate whether this can be used to enrich the cliques with person-to-person relations.

AUTHOR CONTRIBUTIONS

APA and ST designed the work and wrote the paper. APA and GM implemented the software to run the experiments. SM and APA developed the datasets for evaluation.

FUNDING

This work has been funded by Fondazione Bruno Kessler.

REFERENCES

Bron, C., and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM* 16: 575–7. doi:10.1145/362342.362367

Corcoglioniti, F., Rospocher, M., and Aprosio, A.P. (2016). A 2-phase frame-based knowledge extraction framework. In *Proc. of ACM Symposium on Applied Computing (SAC'16)*. Pisa, Italy.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P.N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, Graz, Austria.

Diesner, J., and Carley, K. (2009). He says, she says. Pat says, Tricia says. How much reference resolution matters for entity extraction, relation extraction, and social network analysis. In *Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on*, 1–8. Ottawa, Canada.

Diesner, J., Evans, C.S., and Kim, J. (2015). Impact of entity disambiguation errors on social network properties. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015*, 81–90. Oxford, UK: University of Oxford.

Downey, R.G., and Fellows, M.R. (1995). Fixed-parameter tractability and completeness II: on completeness for W[1]. *Theoretical Computer Science*, 141: 109–31. doi:10.1016/0304-3975(94)00097-3

Elson, D.K., Dames, N., and McKeown, K.R. (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, 138–147. Stroudsburg, PA, USA: Association for Computational Linguistics.

Finkel, J.R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of ACL '05*, 363–370. Ann Arbor, USA: Association for Computational Linguistics.

Fokkens, A., Soroa, A., Beloki, Z., Ockeloen, N., Rigau, G., van Hage, W.R., et al. (2014). Naf and gaf: linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, 9–16. Reykjavik, Iceland.

Grabowicz, P.A., Aiello, L.M., Eguiluz, V.M., and Jaimes, A. (2013). Distinguishing topical and social groups based on common identity and bond theory. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13*, 627–636. New York, NY, USA: ACM.

Gregori, E., Lenzini, L., and Orsini, C. (2011). k-clique communities in the internet as-level topology graph. In *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on*, 134–139. Minneapolis, USA.

Hasegawa, T., Sekine, S., and Grishman, R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics, ACL '04*, Stroudsburg, PA, USA: Association for Computational Linguistics.

Henderson, K., and Eliassi-Rad, T. (2009). Applying latent dirichlet allocation to group discovery in large graphs. In *Proceedings of the 2009 ACM Symposium on Applied Computing, SAC '09*, 1456–1461. New York, NY, USA: ACM.

Jin, L., Chen, Y., Wang, T., Hui, P., and Vasilakos, A. (2013). Understanding user behavior in online social networks: a survey. *Communications Magazine IEEE* 51: 144–50. doi:10.1109/MCOM.2013.6588663

Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., et al. (2009). Media meets semantic web – how the BBC Uses DBpedia and linked data to make connections. In *The Semantic Web: Research and Applications: 6th European Semantic Web Conference, ESWC 2009*, 723–737. Heraklion, Crete, Greece: Springer Berlin Heidelberg.

Koper, R. (2004). Use of the semantic web to solve some basic problems in education: increase flexible, distributed lifelong learning, decrease teacher's workload. *Journal of Interactive Media in Education* 2004, 1–23. doi:10.5334/2004-6-koper

Kuter, U., and Golbeck, J. (2009). Semantic web service composition in social environments. In *The Semantic Web – ISWC 2009: 8th International Semantic Web Conference, ISWC 2009*, 344–358. Chantilly, VA, USA: Springer Berlin Heidelberg.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., et al. (2015). DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal* 6: 167–195. doi:10.3233/SW-140134

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60. Baltimore, USA.

Mcauley, J., and Leskovec, J. (2014). Discovering social circles in ego networks. *ACM Transactions on Knowledge Discovery from Data* 8: 28. doi:10.1145/2556612

Melamed, I.D., and Resnik, P. (2000). Tagger evaluation given hierarchical tag sets. *Computers and the Humanities* 34: 79–84. doi:10.1023/A:1002402902356

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics* 7: 767–73. doi:10.1016/j.joi.2013.06.006

Mitchell, A., Strassel, S., Przybocki, M., Davis, J., Doddington, G., Grishman, R., et al. (2002). *ACE-2 Version 1.0. LDC2003T11*. Philadelphia, USA: Linguistic Data Consortium.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Moretti, G., Sprugnoli, R., Menini, S., and Tonelli, S. (2016). ALCIDE: extracting and visualising content from large document collections to support humanities studies. *Knowledge Based Systems* 111: 100–12. doi:10.1016/j.knosys.2016.08.003

Özgür, A., Cetin, B., and Bingol, H. (2008). Co-occurrence network of Reuters news. *International Journal of Modern Physics C* 19: 689–702. doi:10.1142/S0129183108012431

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435: 814–8. doi:10.1038/nature03607

Palmero Aprosio, A., and Giuliano, C. (2016). The Wiki Machine: an open source software for entity linking and enrichment. *FBK Technical report*.

Palmero Aprosio, A., Giuliano, C., and Lavelli, A. (2013a). Automatic expansion of dbpedia exploiting wikipedia cross-language information. In *ESWC, Volume 7882 of Lecture Notes in Computer Science*, Edited by P. Cimiano, V. Corcho, L. Presutti, L. Hollink, and S. Rudolph, 397–411. Berlin, Heidelberg: Springer.

Palmero Aprosio, A., Giuliano, C., and Lavelli, A. (2013b). Automatic mapping of Wikipedia templates for fast deployment of localised DBpedia datasets. In *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies, i-Know '13*, 1–8. New York, NY, USA: ACM.

Palmero Aprosio, A., and Moretti, G. (2016). Italy goes to Stanford: a collection of CoreNLP modules for Italian. *ArXiv e-prints*.

Pei, J., Jiang, D., and Zhang, A. (2005). On mining cross-graph quasi-cliques. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05*, 228–238. New York, NY, USA: ACM.

Rizzo, G., and Troncy, R. (2012). Nerd: a framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL '12*, 73–76. Stroudsburg, PA, USA: Association for Computational Linguistics.

Sudhahar, S., Veltri, G.A., and Cristianini, N. (2015). Automated analysis of the US presidential elections using big data and network analysis. *Big Data and Society* 2:1–28. doi:10.1177/2053951715572916

Veling, A., and Van Der Weerd, P. (1999). Conceptual grouping in word co-occurrence networks. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence – Volume 2, IJCAI'99*, 694–699. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.