NELasso: Group-Sparse Modeling for Characterizing Relations among Named Entities in News Articles

Amara Tariq, *Student Member, IEEE,* Asim Karim, *Member, IEEE,* Hassan Foroosh, *Senior Member, IEEE*

Abstract—Named entities such as people, locations, and organizations play a vital role in characterizing online content. They often reflect information of interest and are frequently used in search queries. Although named entities can be detected reliably from textual content, extracting relations among them is more challenging, yet useful in various applications (e.g. news recommending systems). In this paper, we present a novel model and system for learning semantic relations among named entities from collections of news articles. We model each named entity occurrence with sparse structured logistic regression, and consider the words (predictors) to be grouped based on background semantics. This sparse group LASSO approach forces the weights of word groups that do not influence the prediction towards zero. The resulting sparse structure is utilized for defining the type and strength of relations. Our unsupervised system yields a named entities' network where each relation is typed, quantified, and characterized in context. These relations are the key to understanding news material over time and customizing newsfeeds for readers. Extensive evaluation of our system on articles from TIME magazine and BBC News shows that the learned relations correlate with static semantic relatedness measures like WLM, and capture the evolving relationships among named entities over time.

Index Terms—Sparse group learning, LASSO, Named entities, Semantic network construction, News understanding

1 Introduction

The Web is a vast collection of information about different concepts, events, and most importantly named entities such as persons, organizations, and locations. Named entities are popular subjects of interest of users, and form common query terms in search engines. A study of blog search confirms that the most popular type of queries are about named entities [1]. News articles, which are similar to blogs, almost always refer to some named entities while describing events, concepts, and opinions. Thus, named entities are an important aspect of information retrieval, recommendation, and personalization. For instance, a named entities' based representation of news articles is utilized for customizing newsfeeds to users in [2].

Besides simply detecting named entities in documents such as news articles, which is a well-studied problem, discovering and understanding relations among named entities can provide additional insights for enhanced retrieval, navigation, and customization tasks. For example, knowing that 'Mitt Romney' and 'Paul Ryan' are related can guide users to additional relevant news articles. Furthermore, knowing the context in which the relation exists helps in understanding the relationship between the entities (for the example relation above, the context is the 2012 U.S. presidential election). Semantic relations between named entities can be found from resources like Wikipedia[1]. However, such relations cannot capture the dynamics of the relationships over time including their changing contexts.

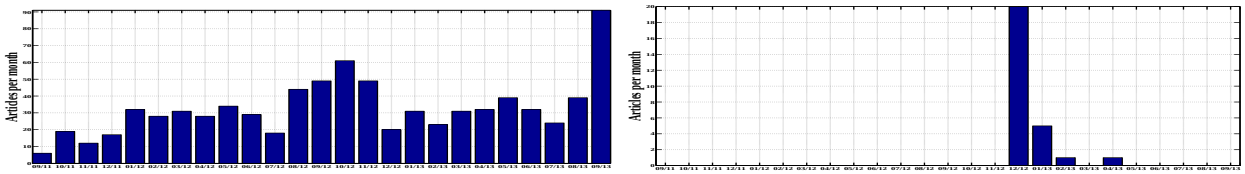
Figure 1 shows the distribution of two named entities over time in news articles published in the TIME magazine. While entities like 'Barack Obama' occur frequently throughout the observed time period (from April 2010 to October 2013), the frequency of another entity 'Adam Lanza' peaks for a short time (corresponding to the event of Sandy Hook elementary school shooting incident in Newtown, Connecticut, USA). In general, the distribution of named entities in news articles varies over time.

Previously, various systems have been proposed for detecting relations between named entities. Some of these systems require external resources such as Wikipedia or Freebase[2] for their operation [3], [4]. Some supervised systems focus on discovering a few pre-defined types of relations [5], [6], [7]. Usually, such systems need an initial seed in the form of pairs of named entities with specific types of relations between them [8], [9]. OpenIE (open information extraction) systems extract relational tuples without pre-specification of a vocabulary, but require the related entities to be mentioned in a single sentence with a certain structure [10], [11], [12]. None of these systems are capable of machine understanding news material through relations among named entities discovered from news articles over time.

In this paper, we present NELasso, an automatic system and model for discovering and characterizing relations among named entities mentioned in collections of news articles. NELasso models named entity occurrences via sparse structured logistic regression. The words that strongly predict the occurrence of a named entity identify its context, while the sparsity inducing learner ensures that less relevant (or noisy) words are removed from the model. We impose a group structure over the words based on background knowledge (e.g., groups based on keywords, topics, and co-occurrence patterns). The relation between two named entities is defined by the common groups of

1. www.wikipedia.com

2. https://www.freebase.com/



(a) Barack Obama (b) Adam Lanza

Fig. 1: Distribution of named entities over time; each bar represents frequency of a named entity during one month

words that strongly predict the two named entities. These word groups also yield the relation *type* (a context descriptor for the relation). We also propose a measure for quantifying the *strength* of relations between named entities.

We evaluate our system extensively on two news datasets, through automatic evaluation, manual/human assessments, and comparisons with baseline systems. The results demonstrate the effectiveness of our system in discovering and characterizing significant relations among named entities. NELasso has the following desirable features: (1) It is completely unsupervised in nature and does not require specification of relation templates or contexts, (2) It allows flexible definitions of relation types through appropriate structures over the words, and (3) It tracks the dynamics of relations over time via their changing strengths and types.

The rest of the paper is structured as follows. Section 2 provides a review of previous work regarding sparse learning, named entity linking, and semantic relation extraction. Section 3 presents the problem setting and the proposed system and model for its solution. We describe the formation of group structure for sparse learning in Section 4. Section 5 discusses the output of our system on real-world datasets. A comprehensive evaluation of our system is presented in Section 6. We conclude our contributions in Section 7.

2 Motivation and Related Work

We motivate our proposed system by discussing related work on sparse learning and named entity relation discovery. The LASSO (least absolute shrinkage and selection operator) learning model was introduced by Tibshirani [13]. This operator emphasizes sparsity in the learned model through an ℓ_1-norm penalty, thus rendering it more interpretable than other models. Variations of this model have been devised to consider inherent structure in the feature set [14], [15], [16]. Usually, structure in feature set is depicted by groups of features that are related in some way and sparse structured learning penalizes groups of features instead of (or in addition to) individual features. Thus, models for sparse structured learning include both ℓ_1- and ℓ_2-norm based penalties to induce sparsity on and within groups of features. Sparse structured learning models have typically been tailored for specific applications (e.g., [17]). Lasso and structured lasso models have been adopted for a range of natural language processing (NLP) and text mining problems, e.g., text categorization, semantic similarity between words, chunking, entity recognition, and dependency parsing [18], [19], [20], [21].

Much work has been done on extracting named entities (persons, organizations, and locations) and semantic relation between words, concepts, and named entities in the semantic web and NLP communities. For example, the task of entity linking in semantic web research aims to link entities mentioned in text to some known database of named entities such as Wikipedia [22], [23], [24]. BabelNet is a large corpus of semantic relations between words from different languages [25], while another system for this task is presented by Dai et al. [26]. Szumlanski et al. present a method to automatically build a large semantic network of concepts using WordNet and Wikipedia [3].

OpenIE (open information extraction) is the process of extracting relational tuples from text without prespecification of vocabulary. Such systems can only discover relations between entities mentioned in a single sentence with a specific structure, such as involving a verb [10], [11]. Schmitz et al. propose OLLIE which eliminates the restriction of the sentence being mediated by a verb [12]. But, OLLIE still requires a seed from the output of ReVerb system [11] and the related entities to be mentioned in a single sentence obeying any of the learned patterns.

Systems for named entity relation extraction are usually supervised or semi-supervised in nature [4], [5], [6], [7], [8], [9]. Such systems require initial seed information in the form of pairs of related named entities. Subsequently, they find other pairs of named entities with the same type of relation between them as in the seed relations. As such, these systems are restricted to discovering relations of a limited variety and may also require external databases for their working. Less work has been reported on unsupervised methods for relation extraction. Rosenfeld et al. propose clustering of named entities based on their context in documents [27], while Hasegawa et al. propose unsupervised relation discovery among named entities appearing in the same sentence [28]. These unsupervised systems impose restrictions on the named entities to be considered for relation discovery and assign only a handful of manually picked labels to discovered relations.

Most of the previously proposed systems are aimed at building a knowledge-base of facts [4], [5], [6], [7], [8], [9], [11], [12]. As such, these systems extract relations having a limited number of linguistic patterns connecting named entities in single sentences. In comparison, our system is aimed at machine understanding of news material through the discovery and characterization of relations among named entities mentioned in news articles. In other words, our *unsupervised* system considers two named entities to be related when they appear in the same context within a collection of news articles, taking into account the distribution of entities over articles in the collection instead of limiting the processing to sentences. Our system also assigns statistical signatures to relations which are useful in news customizing and search/browsing applications. In short, the scope and nature of our system is largely different from that of other systems presented in the literature.

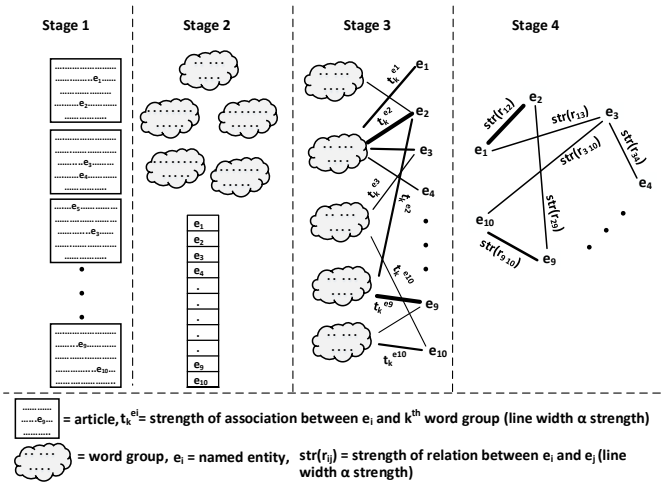


Fig. 2: Overview of NELasso: Stage 1 = news articles collection, Stage 2 = named entities extraction (Section 4.1.2), word group formation (Section 4.2), Stage 3 = learn sparse group model associating named entities and word groups (Section 3.2.2), Stage 4 = quantify semantic relations between named entities and build network (Section 3.2.3)

3 NELasso: System and Model

In this section, we present an overview of our system and the details of sparse group lasso model underlying NELasso.

3.1 System Overview

Figure 2 displays the overall working of our proposed system, NELasso. The input to the system is a collection of news articles (Stage 1) and the output is a semantic network of named entities mentioned in those articles (Stage 4). The system preprocesses the articles, extracts the named entities appearing in them, and builds groups of related words used in the articles (Stage 2). It then employs sparse learning to determine the association between named entities and word groups (Stage 3). The system characterizes and quantifies the relations between named entities and builds a semantic network from this information (Stage 4).

3.2 Modeling Named Entity Relations

NELasso is based on a sparse group lasso model of named entities appearing in text documents. It is an automatic and unsupervised model for discovering significant relations of different types that can be specified in a flexible manner by imposing a group structure over the words in the collection. In this section, we formalize our model for named entity relation discovery and quantification.

3.2.1 Notation and Problem Setting

We are given a collection of documents D (e.g., news articles) defined over the vocabulary set V of unique words. Let M and N be the numbers of articles and words, respectively, in D and V. We know the set of named entities E mentioned in the collection of articles D (e.g., by using a named entity recognizer), and a specific article $d \in D$ can contain zero or more named entities from E. The words in the vocabulary set are partitioned into $K > 1$ subsets. These groups of words encapsulate additional semantic knowledge of words in different contexts (e.g., topics) within the collection and help characterization of relations between named entities.

Given the above setting, we seek relations $r_{ij} = rel(e_i, e_j)$ between entities $e_i \in E$ and $e_j \in E$ $(i \neq j)$ that are significant in D. Each relation is characterized by its *type*, $type(r_{ij})$, and its *strength*, $str(r_{ij})$, where the *type* qualifies the context of the relation and the *strength* quantifies it. These results are obtained in an unsupervised manner.

Intuitively, a relation r_{ij} is likely to exist in D when both entities e_i and e_j are mentioned in the same context (e.g., event) in D. We formulate this intuition to discover and characterize relations between named entities as discussed in the following subsections.

3.2.2 Sparse Structured Modeling of Named Entities

The occurrence of a named entity in news articles depends on the context (topic, event, story, etc.) of the articles in which it is mentioned, and the context is specified by the words used in those articles. We use this idea to model the named entity's occurrence as a classification problem, where the words appearing in articles serve as predictors and the occurrence of the named entity as the target or response. A separate model is learned for each named entity in E.

Since not every word plays a significant role in predicting every entity, we adopt a sparsity inducing approach by introducing an ℓ_1-norm of coefficients as a penalty to the standard classification objective function. We also impose a penalty on all the coefficients of words from each group. This penalty, which is also added to the objective function, is the ℓ_2-norm of the coefficients of each word group. The latter penalty tries to eliminate entire groups of words from the model, thus further enhancing sparsity and interpretability of the model, especially when groups carry contextual semantics.

Consider a single named entity $e \in E$. Then, the sparse group lasso logistic regression model for the named entity e is given by the following minimization problem:

$$\min_{\mathbf{x}} [\sum_{m=1}^{M_e} \ln(1 + \exp(-y_m(\mathbf{x}^{\mathrm{T}}\mathbf{d}_m + c))) + \lambda_1 ||\mathbf{x}||_1 ... \\ ... + \lambda_2 \sum_{k=1}^{K} \phi_k ||\mathbf{x}_k||_2] \quad (1)$$

Here, $y_m \in \{-1, +1\}$ indicates whether mth article mentions the entity ($y_m = +1$) or not ($y_m = -1$). M_e is the number of articles used in training. In practice, we prefer to have articles that mention and do not mention the entity in almost equal proportions in the training set; thus, the training set for $e \in E$ includes all articles that mention e and the same number of randomly picked articles that do not mention e. Therefore, $M_e \leq M$ in general. The vector $\mathbf{d}_m \in \Re^N$ represents the mth article in bag-of-words format. The vector $\mathbf{x} \in \Re^N$ contains the learned coefficients corresponding to the words in $\mathbf{d}_m$.

The model assumes a group structure among words such that the coefficient vector $\mathbf{x}$ consists of K non-overlapping groups of coefficients $\mathbf{x}_k$. The term ϕ_k assigns an additional weight/penalty to the kth group of coefficients. These terms

can be selected empirically, but in most cases in practice (including our experiments), they can be set to one.

There are two regularization parameters or terms in the ℓ_1/ℓ_2 regularized logistic regression model. The first term λ_1 rewards the selection of fewer words, while the second term λ_2 enforces sparsity on the group structure of the words – it rewards selection of as few groups as possible from the available groups of words. The sparse group lasso model can be solved efficiently by the implementation provided in the SLEP package[3]. This implementation also finds the optimal values of the regularization parameters automatically.

3.2.3 Semantic Relations Among Named Entities

A sparse group-structured model, i.e., the vector $\mathbf{x}$ containing the coefficients corresponding to the words in the vocabulary set, is estimated for each named entity in E. This information, together with how these coefficients exist across groups, is used to establish relations among named entities. With this information, we also define the *type* as well as the *strength* of each relation.

A word provides positive evidence for a named entity e if the value of the corresponding coefficient in the entity's prediction model is greater than zero. The evidence provided by words in the kth group for entity e, denoted by t_k^e, is estimated by summing up entries x_n of the coefficient vector $\mathbf{x}$ such that this nth word/coefficient belongs to group k and $x_n > 0$. We say that this evidence is significant when it is greater than a threshold, i.e., $t_k^e \geq \gamma$ where $\gamma \geq 0$ is a selection threshold. The value of γ decides the amount of positive evidence a group of words needs to provide for a named entity for it to be considered as a contender for establishment of semantic relations.

Consider a relation r_{ij} between entities e_i and e_j. The *types* of this relation are given by the groups of words that provide significant positive evidence for both entities e_i and e_j. For example, if the group defined by the keyword 'Election' provides significant positive evidence for named entities 'Mitt Romney' and 'Paul Ryan' then the *type* of the relation between these named entities is 'Election'. In general, one or more *types* can characterize a relation. If no groups provide significant positive evidence for both of the entities, then no relation exists between them. The *strength* of relation r_{ij} of *type* k is defined as $str_k(r_{ij}) = t_k^{e_i} \times t_k^{e_j}$.

Definition 1. (Relation r_{ij}) A relation r_{ij} of *type* k exists between entities e_i and e_j in D when both $t_k^{e_i}$ and $t_k^{e_j}$ are greater than the selection threshold $\gamma \geq 0$. Here, t_k^e is the sum of the positive coefficients in the kth group in the sparse group logistic regression model for the entity e. The *strength* of r_{ij} is defined as $str_k(r_{ij}) = t_k^{e_i} \times t_k^{e_j}$.

Note that semantic relations between named entities are not dependent on their association with individual words but with groups of similar words. Each group is the lexicon for one particular context (e.g., topic). This ensures that relations between named entities are not ignored because of the use of different individual words. Rather, they are estimated based on whether or not both named entities relate to the lexicon of the same context, thus enabling the system to identify more complex relationships.

3. http://www.public.asu.edu/ jye02/Software/SLEP/

4 NELasso: Vocabulary and Group Structure

In this section, we present the preprocessing steps and group structure formation methods in NELasso.

4.1 Preprocessing of News Articles

The system is given a news articles collection along with its metadata, e.g., article publication date and keywords. This collection is processed to form a vocabulary set and to identify the named entities mentioned in the articles.

4.1.1 Vocabulary Set

TreeTagger[4] is used to tokenize, lemmatize and part-of-speech tag news articles' text. Frequently occurring nouns, verbs, and adjectives are retained to form the vocabulary set V. Each article is represented as a vector $\mathbf{d}$ of length V where $N = |V|$. Each element of $\mathbf{d}$ records the number of times the corresponding word set V occurs in an article. This is standard bag-of-words representation for text documents.

4.1.2 Named Entities

There are many techniques for identifying named entities reliably from text documents. Our system uses the Stanford named entity recognizer (NER)[5] trained over MUC named entity corpora that identifies 7 different classes of entities, i.e., Person, Organization, Location, Time, Percent, Money, and Date. The system retains only named entities of types Person, Organization, and Location as these are the most interesting and important entities mentioned in news articles.

4.2 Group Structure of Vocabulary Set

The formation of word groups is a key step in our system. We seek $K > 1$ groups of words of the vocabulary set such that all words in a particular group capture a specific context in news articles (e.g., topic, event, keyword(s)). We exploit multiple sources of information to form these groups, as explained in the following subsections.

4.2.1 Co-occurrence-based Word Groups

The co-occurrence of words in articles can be used to form groups of related words. Typically, each article discusses a specific topic or news story. The presence of a word in an article indicates its relationship with the topic or story of the article. Two articles that contain similar words are likely discussing the same topic or news story. Thus, co-occurrence of vocabulary words in the same set of articles is an important clue for forming word groups.

To find co-occurrence-based word groups, we represent each word w_j in the vocabulary set by a vector $\mathbf{v}_j$ of length M where M is the number of articles in the collection. The ith element v_{ji} of this vector indicates presence (1) or absence (0) of the word w_j in the ith article. Our system employs agglomerative hierarchical clustering of these vectors to find groups of related words. The cosine similarity is adopted for comparing vectors; the single-link merge operator is used; and a constraint is imposed to restrict cluster size

4. http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
5. http://nlp.stanford.edu/software/CRF-NER.shtml

to τ or less. Iteratively, any cluster larger than τ is further divided. This procedure results in a finite number of non-overlapping subsets V_k ($k = 1, \ldots, K$) of the vocabulary set V such that $\forall k, |V_k| \leq \tau$ and $\forall l, t,\ V_l \cap V_t = \emptyset$. Threshold τ determines the maximum allowed size of the word groups, and hence the number of word groups formed.

4.2.2 Keyword-based Word Groups

News websites often assign one or more keywords to each article which characterize its topical context and help the reader navigate easily to other articles discussing the same topic. Words appearing in articles having a certain keyword are obviously indicative of the topic or news event represented by that keyword. Thus, keywords can aid the process of identifying groups of words associated with a topic.

We need to estimate the importance of each vocabulary word in identifying a particular topic represented by a specific keyword. This scenario is similar to the term-to-topic relatedness concept introduced in [29], [30]. Relationships between words, i.e., term-to-term relationships are commonly used in many natural language processing tasks. However, relationships between words and topics, i.e., term-to-topic relationships are more useful in cases where context/topic is already known. In our setting, topic or context is specified in the form of keywords assigned to news article.

The relatedness of a word in the vocabulary set with a context defined by a keyword can be quantified by its discriminative term weight (DTW). The DTW for vocabulary word w_j given context/keyword C_k is defined as

$$dtw(w_j, C_k) = \frac{p(w_j|C_k)}{p(w_j|C_k^{'})} \tag{2}$$

Here, $p(w_j|C_k)$ is the probability of word w_j in news articles belonging to keyword C_k while $p(w_j|C_k^{'})$ is the probability of word w_j belonging to news articles of every keyword other than C_k [29], [30]. To estimate these probabilities, we assume a document model in which each vocabulary word follows the Bernoulli distribution, i.e., the word either occurs or does not occur in articles of a given keyword.

Once the DTW of all words with respect to all keywords have been calculated, each word is associated to the keyword for which it has the highest DTW. Let key_j denotes the keyword to which word w_j has been assigned, then

$$key_j = \arg\max_k dtw(w_j, C_k) \tag{3}$$

Thus, each vocabulary word is assigned to one keyword. If V_k represents a subset of the vocabulary set V consisting of all words assigned to the keyword C_k, then $\forall l, t\ V_l \cap V_t = \emptyset$, i.e., vocabulary set is divided into non-overlapping subsets such that there is one subset for each keyword.

Oftentimes, the distribution of articles among keywords can be extremely uneven. Some keywords, such as 'World' (in TIME dataset), are too general and are assigned to a large number of articles covering many different topics. Thus, the word groups for such keywords are very large. To address this issue, we further divide word groups V_k with $|V_k| > \tau$ into smaller word groups using co-occurrence of words, as described in Section 4.2.1. For example, the word group corresponding to the keyword 'World' may now be divided into subgroups 'World1', 'World2', etc.; each subgroup corresponding to one news story covered by articles of keyword 'World'. In this process, the threshold τ determines the maximum allowed size of the word groups and affects the number of word groups formed, as it does for co-occurrence based word groups.

4.2.3 Topic-based Word Groups

Topic modeling is a powerful tool for document collection understanding. In a topic model, each document is considered as a mixture of topics and each word in the vocabulary set has a distribution over the topics. This distribution quantifies the relation of a word with all the topics. LDA (latent Dirichlet allocation) is the most famous generative topic model today [31]. Since our system requires identification of word groups belonging to certain topics, we propose a similar generative model for our news article collection.

- choose $\theta_d \sim \text{Dir}(\alpha)$
- for each of the S words w_j in article
  - choose topic $C_k \sim \text{Multinomial}(\theta_d)$
  - choose a word w_j from $p(w_j|C_k, \beta)$, a Multinomial probability conditioned on topic C_k

θ_d is a K-dimensional Dirichlet random variable, where K is the number of underlying topics that generated the article collection. K needs to be fixed before starting the estimation process. β is a fixed quantity to be estimated. S is the length of the article which is assumed as a fixed number. $p(C_k|\theta_d)$ is the probability of the topic C_k given the news article d.

The estimated distribution $p(w_j|C_k, \beta)$ captures the association of each vocabulary word w_j with each underlying topic C_k. We take the set of underlying topics of the article collection as the basis for word group formation. The system uses a threshold ε on the value of $p(w_j|C_k, \beta)$ to decide whether or not the word w_j belongs to the topic C_k. Thus, there are as many word groups as the number of topics (K). The group corresponding to topic C_k contains all the words with reasonably high conditional probability given C_k.

In general, this method forms overlapping word groups. Each word group V_k is a subset of vocabulary set V such that $V_l \cap V_t \neq \emptyset$. For this method, our system repeats words appearing in multiple word groups in the vector $\mathbf{d}_m$ for the mth article.

5 System Output

In this section, we discuss the results of our system for discovering and characterizing relations among named entities on two news articles' collections.

5.1 Datasets

We use two datasets in our experiments. The first dataset contains news articles collected specifically for this work from the TIME magazine website[6]. This dataset contains $19,841$ news articles with their publication dates from early 2007 to late 2013. The highest concentration of articles is from June 2010 to September 2013. Therefore, we focus on this time period in our experiments. To conduct a time-evolving analysis of the articles, we divide this time period

6. www.time.com

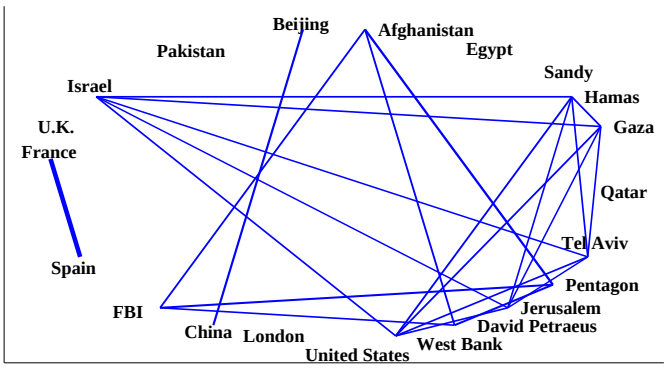


Fig. 3: Semantic network of named entities for Nov-Dec 2012 (TIME dataset)

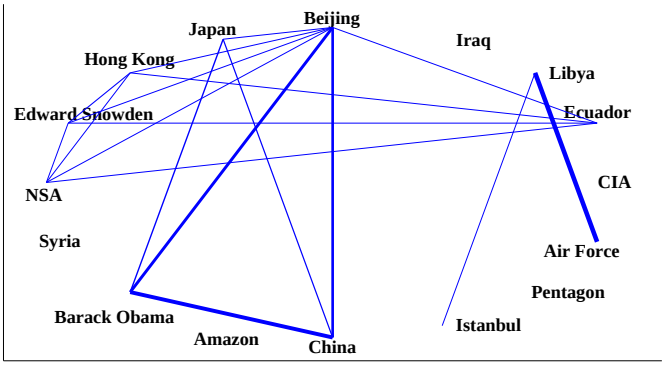


Fig. 4: Semantic network of named entities for Jun-Jul 2013 (TIME dataset)

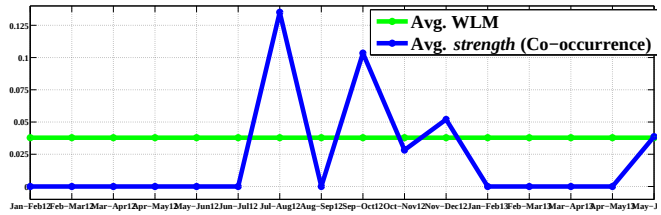
into smaller time slots, generally of one month duration, and run our system for each time slot. There are $6,350$ words in the vocabulary set and about $1,100$ named entities of interest in this dataset. Not all entities of interest are mentioned in articles of each time slot. Furthermore, only entities that occur frequently in a time slot are considered for relationship building. There are 719 unique keywords associated with articles in this dataset.

The second dataset is a previously available collection of articles from BBC UK[7]. This dataset contains $3,352$ articles with a vocabulary set of $5,995$ words. We retained 398 frequently occurring named entities in this dataset for analysis. This dataset does not have publication date and keywords associated with articles. Therefore, we tested only two out of the three methods of word group formation discussed earlier (excluding keyword-based word groups). Since publication date information is unavailable, we tested our system over 10 random subsets of the dataset, of 500 articles each.

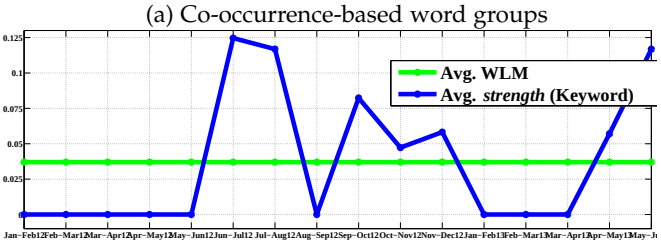
5.2 Network of Named Entities

The relations among named entities in a given time period can be presented visually as a semantic network. Figures 3 and 4 are two sample semantic networks generated by our system for two different time slots of the TIME dataset. An edge between two named entities indicates a relation between those entities. The thickness of an edge indicates the *strength* of the strongest *type* of relation between the entities. Entities such as 'Gaza', 'Hamas', 'West Bank', 'Tel Aviv', 'Jerusalem', and 'Israel' are connected to each other with thick edges in Figure 3. This network corresponds to the time when news articles were being published about Operation Pillar of Defense which involved these entities. In Figure 4, entities such as 'Edward Snowden', 'NSA', 'Ecuador', and 'Hong Kong' are connected to each other as this network corresponds to the time when the NSA leaks story broke out. The networks of named entities produced by our system provide an intuitive understanding of news stories and named entities discussed in a given time period.

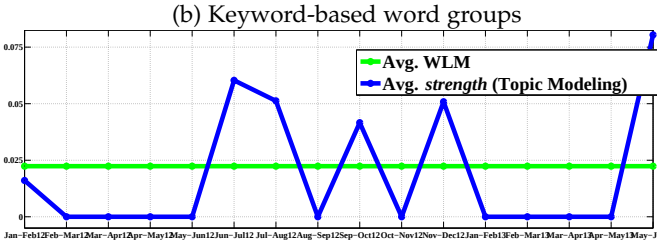
7. http://homepages.inf.ed.ac.uk/s0677528/data.html



(a) Co-occurrence-based word groups



(b) Keyword-based word groups



(c) Topic-based word groups

Fig. 5: Variation of average *strength* and WLM of relations over time (TIME dataset). Relations among the following named entities exist in each time period: {Syria, Bashar Asad, Cairo, Damascus, Jerusalem, Hamas, Gaza, Israel Egypt, Benghazi, Hillary Clinton}; x-axis: Time period (months). y-axis: Mean WLM and relation *strength*

5.3 News Events

In the semantic networks built by our system, we can identify cliques of related named entities. A clique in a network is a group of named entities in which every named entity is related to every other named entity in the group and all relations are of the same *type*. These cliques typically correspond to major events in the news articles' dataset, and provide a summary at a glance of named entities involved in the events. Table 1 gives some cliques identified by our system along with the time period in which they occur and their *type*. For keyword-based word groups, the keyword provides a label for the relation *type*. For co-occurrence- and topic-based word groups, relation *type* has been indicated by a few top words of the word group responsible for the connection among the named entities. In any case, the relation *type* points to the news story in which the named entities of the corresponding clique play important roles.

Named Entities	Time Period	Relation *type*		
		Keyword	Co-occurrence	Topic
Colorado, James Holmes, Aurora	Jul-Aug,2012	Crime	Aurora, Shooting, Theater	Kill, Shooting, Colorado
South Korea, Pyongyang, North Korea, Kim Jong Il	Dec,2011-Jan,2012	North Korea	North, Korean, Korea Imperial, Successor	North, Korea, Leader Military , Dictator
Israel, Hamas, Tel Aviv, Gaza, Jerusalem, West Bank	Nov-Dec,2012	Israel	Gaza, Hamas, Radical Israel, Occupation	Israel, Palestinian, Fire Rocket, Refugee, Gaza

TABLE 1: Example cliques discovered by our system (TIME dataset); each clique corresponds to a distinct news event indicated by the *type* of the relation

Related Named Entities	Time Period	Associated News Story
Mitt Romney - South Carolina (0.04)	Jan-Feb, 2012	South Carolina Republican Primary: Jan 21, 2012
Mitt Romney - Florida (0.03)	Jan-Feb , 2012	Florida Republican Primary: Jan 31, 2012
Mitt Romney - Arizona (0.03)	Feb-Mar, 2012	Arizona Republican Primary: Feb 28, 2012
Mitt Romney - Ohio (0.05)	Feb-Mar, 2012	Ohio Republican Primary: Mar 06, 2012
Mitt Romney - Illinois (0.07)	Mar-Apr, 2012	Illinois Republican Primary: Mar 20, 2012
Mitt Romney - Paul Ryan (0.11)	Aug-Sep, 2012	Mitt Romney announced Paul Ryan as his running mate on August 11, 2012
Mitt Romney - Tampa (0.06)	Aug-Sep, 2012	Mitt Romney formally accepted Republican Party nomination on August 30, 2012 in Tampa, Florida.

TABLE 2: Example of a named entity involved in different relations over time (TIME dataset). The *strength* of the relation using keyword based word groups is shown in parenthesis

Related Named Entities	Time Period	*type* #1	*type* #2
Spain - U.K.	Oct-Nov, 2012	BBC, Live, Set, International, European	Economics, Rise, Growth, Spending, Crisis
Mitt Romney - White House	Oct-Nov, 2012	President, Presidential , Debate, Obama, Romney	Election, Candidate, Vote, Poll, Race
Iran - Russia	Mar-Apr, 2013	Diplomat, Negotiation, Sanction, Suspension	Aggression, Ballistic, Firing, Hostile, Target
Turkey - Istanbul	Jun-Jul, 2013	Police, Protest, Street, Night, President	War, Syria, Rebel, Assad, Regime

TABLE 3: Examples of relations with more than one relation *type* in one time period (TIME dataset)

5.4 Dynamics of Relations

The output of our system makes it easy to understand the dynamics of relations among named entities over time. Figure 5 shows the variation of average *strength* of relations between all pairs of entities among a selected set of named entities over different time periods in TIME dataset. The set of named entities (given in the figure's caption) is selected such that relations between all pairs of these entities exist in all time periods. These graphs (one each for co-occurrence-, keyword-, and topic-based word groups) show that the average *strength* (blue line) varies greatly over time for the same set of relations. These graphs also show that the average WLM (Wikipedia link-based measure) across all pairs of entities (green line) remains constant over time as WLM is a static measure of relation strength derived from Wikipedia (WLM is discussed in detail in the next section).

The temporal variation in average relation *strength* can be associated with the popularity of news stories involving the selected named entities. The blue lines for all three types of word groups have distinct peaks in July 2012, corresponding to the news story about Damascus bombing involving named entities 'Syria', 'Bashar Asad', 'Damascus', etc. Peaks observed in September 2012, correspond to the Benghazi attack and its aftermath involving discussion on entities such as 'Damascus', and 'Hillary Clinton'. News story of Operation Pillar of Defense involving entities 'Israel', 'Hamas', 'Gaza', 'Egypt', etc., corresponds to the peaks observed in November 2012. Peak in May 2013 correspond to a rare interview of Bashar Asad involving entities 'Syria' and 'Israel'. Our system successfully captures the evolutionary nature of named entities relation in news material.

Our system discovers various relations of 'Mitt Romeny' with other named entities over time (Table 2), correlating with the occurrence of certain news events. Thus, our system can track an entity over time, discovering its relations to specific events or news stories.

Table 3 shows pairs of named entities which are related to each other with more than one relation *types* at the same time. Each relation *type* hints at some news story involving both entities. Our system is flexible enough to deal with the complexity of news material based named entities' relations whereas static relation measures, e.g., WLM, fail to do so.

The outputs of NELasso highlight its suitability for news material understanding, and this is the primary purpose of this system. Previously proposed systems do not possess such a capability and have a different goal altogether, i.e., construction of databases of facts [8], [9], [10], [11], [12].

6 System Evaluation

In this section, we present a comprehensive evaluation of our system. The evaluation includes quality assessment through automatic methods and human judges, comparisons with two baseline methods, sensitivity analysis of system's output with varying parameters, and scalability analysis of the system.

The main output of our system is the semantic network of named entities for a given time period of interest. We define and use in our evaluations two properties of such networks: average degree or connectivity of the network and average *strength* of the relations in the network. If Σ and Υ are the number of edges and nodes, respectively, in the network then its average connectivity is defined as

$$Connectivity = \frac{2 \times \Sigma}{\Upsilon} \quad (4)$$

Our system assigns *strength* to each discovered relation, i.e., $str_k(r_{ij})$ is the strength for relation r_{ij} of *type* k between entities e_i and e_j. This value defines the thickness of the edges

in the network, and the average *strength* of the network is given by the average of these values for the network.

There are two ways of evaluating the system's output: (a) to verify the relations found through an independent source, and (b) to verify that the relations found are helpful in search and retrieval scenarios (a key application of our system). We consider both options in our evaluations and present an automatic method each for option (a) and option (b). Besides automatic evaluation, we also conduct human assessment of the system's output and compare performance with two baseline methods.

The rest of this section is organized as follows. We present our automatic evaluation methods in Sections 6.1 and 6.2. The results of the automatic evaluations are discussed in Section 6.3. We report on the design and the results of a human evaluation study of our system's output in Section 6.4. We introduce two baseline methods and compare our systems output with them in Sections 6.5 and 6.6. We study the effect of system parameters on the quality of the system's output in Section 6.7. Finally, we discuss the time complexity and scalability of our system in Section 6.8.

6.1 Wikipedia-based Evaluation

Our first automatic method of evaluation judges the quality of relations found by our system through an independent source. There are several resources that contain information about relations between named entities. Most significant of such resources is Wikipedia, which is an extensive and highly organized database of information that can be exploited for extracting and characterizing relations among named entities [3], [32].

Specifically, Milne et al. propose a measure, called Wikipedia link-based measure (WLM), to quantify the relatedness between articles a and b based on their inward and outward links in Wikipedia [32]. Each outward link from a and b is assigned a weight given by

$$w(\textit{source} \rightarrow \textit{target}) = \log(\frac{|W|}{|T|}) \text{ if } \textit{source} \in T, 0 \text{ otherwise} \tag{5}$$

where *source* can be a or b and *target* can be other articles in Wikipedia. T is the set of all articles that link to *target* and W is te set of all articles in Wikipedia. The relatedness between a and b based on outward links is defined by the angle between their respective vectors containing weights of their common outward links.

The relatedness between a and b based on inward links is estimated as

$$sr(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))} \tag{6}$$

Here, A and B are sets of all articles that link to a and b, respectively. The final relatedness between a and b is the average relatedness based on outward and inward links. In our setting, a and b correspond to entities e_i and e_j.

We calculate the WLM of a pair of named entities ($wlm(e_i, e_j)$), as explained above, if e_i and e_j are determined to be related by our system. The higher the $wlm(e_i, e_j)$ the stronger is the verification of the relation $rel(e_i, e_j)$ identified by our system through a completely independent source, i.e., Wikipedia. Therefore, we report the mean of $wlm(e_i, e_j)$ for all pairs e_i and e_j identified to be related by our system in a given time period as an evaluation measure of the semantic network of entities built for that time period.

It is worth emphasizing that there are certain advantages to relation extraction though our system over Wikipedia based relation identification. Our system assigns a *type* to each relation and allows the relation between two entities to change its *strength* or *type* or both over time. For instance, entities 'Mitt Romney' and 'Barack Obama' are mentioned frequently in many time periods but relate to each other through relations of varying *type* and *strength* in different time periods. On the other hand, WLM is static over time and provides no clue about the type of relation between two entities. Therefore, this evaluation measure is not meant to judge the quality of relation *type* assigned by our system.

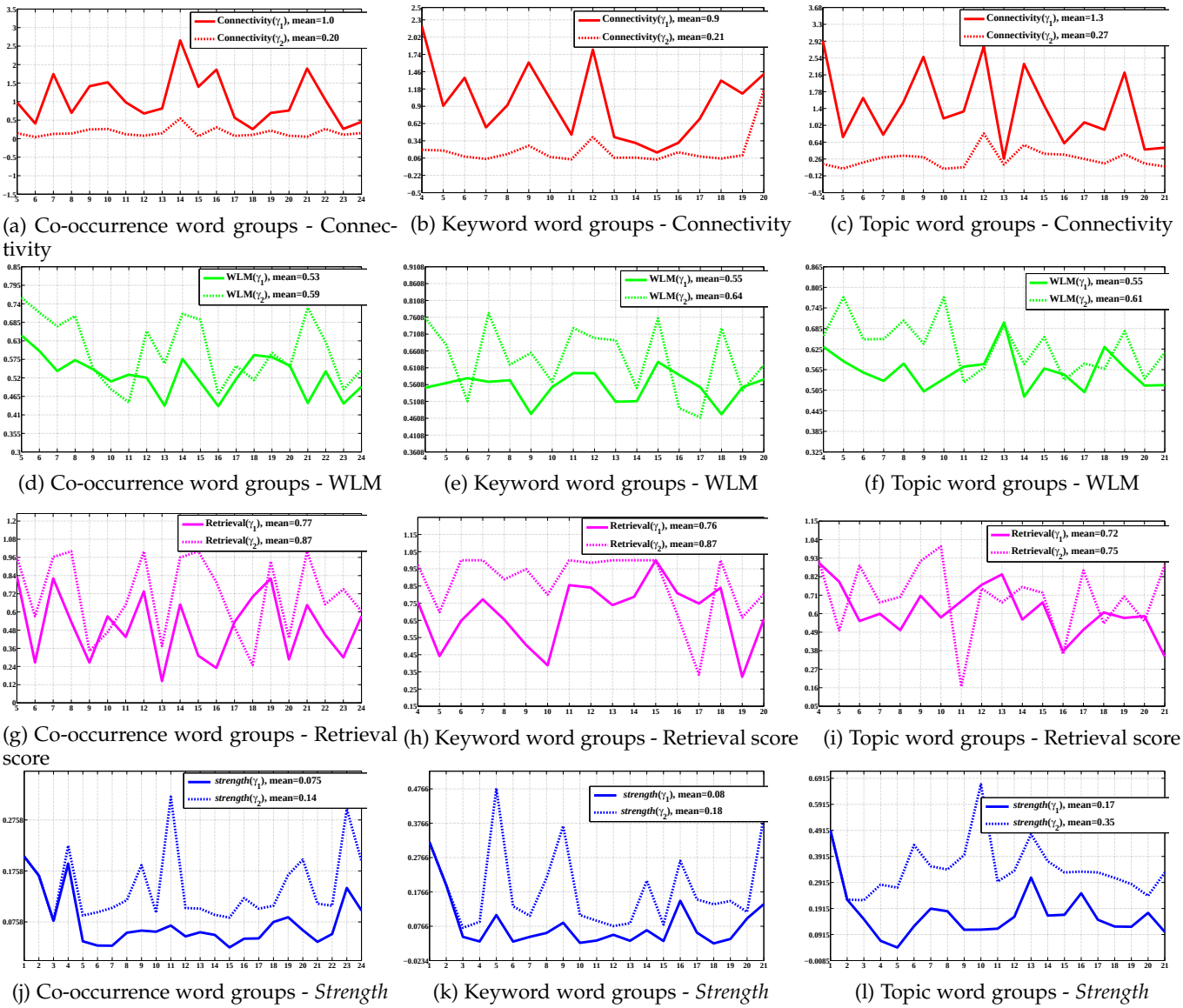
6.2 Retrieval-based Evaluation

Our second automatic method of evaluation judges the usefulness of the relations identified by our system in a retrieval and navigation scenario. The output of our system is a list of relations where each relation r_{ij} between named entities e_i and e_j has a *type* and a *strength* based on some word group, say k. The word group characterizes the context of each relation. If a user is reading a news article that mentions entity e_i in the same context as a relation $rel(e_i, e_j)$, then she should be suggested to read other articles matching the same context and mentioning entity e_j. As explained in Section 1, users reading about 'Mitt Romney' in articles related to 'Election' should be suggested to read about 'Paul Ryan' in other articles of the same context (i.e., 'Election').

The context of a relation needs to be quantified to implement such a recommendation system. We do this by proposing the statistical signature vector ψ^k of length N. Three forms of this vector are developed, one for each type of word group formation explored in our system. If ψ^k_i is the ith entry of vector ψ^k corresponding to $w_i \in V_k$, then ψ^k_i is equal to either (a) sum of term frequency-inverse document frequency (*tfidf*) of w_i of all documents in the given time period for co-occurrence based group formation, or (b) $dtw(w_i, C_k)$ in the given time period for keyword-based group formation, or (c) $p(w_i|C_k, \beta)$ in the given time period for topic-based group formation. All other entries of ψ^k are set to zero.

Using the above quantification of contexts, this evaluation method builds two lists of articles, l_i and l_j, from the given time period for each relation $rel(e_i, e_j)$ of *type* k identified in that time period. Articles in list l_i mention named entity e_i and match context ψ^k, whereas articles in list l_j mention named entity e_j and match context ψ^k. The match between a context and an article is determined by thresholding the cosine similarity between vector ψ^k of the context and the bag-of-words representation vector of the article. We compute the percentage overlap between the two lists as an evaluation measure, called retrieval score, for the relation $rel(e_i, e_j)$ with *type* k. A higher overlap indicates that the topics of discussion regarding the two entities in the given relation are largely the same. Therefore, when a user who is reading about entity e_i is recommended articles about e_j discussed in the same context, she will find the recommendations highly relevant.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2632117, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Co-occurrence word groups - Connectivity (b) Keyword word groups - Connectivity (c) Topic word groups - Connectivity

(d) Co-occurrence word groups - WLM (e) Keyword word groups - WLM (f) Topic word groups - WLM

(g) Co-occurrence word groups - Retrieval score (h) Keyword word groups - Retrieval score (i) Topic word groups - Retrieval score

(j) Co-occurrence word groups - *Strength* (k) Keyword word groups - *Strength* (l) Topic word groups - *Strength*

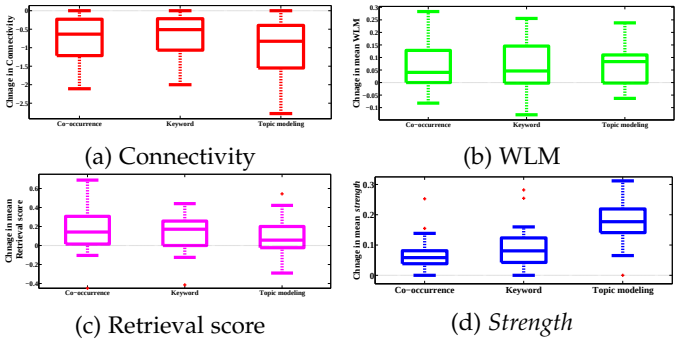
Fig. 6: Effect of threshold γ on evaluation measures on TIME dataset; x-axis: Time period (month), y-axis: Evaluation measure (connectivity, mean WLM, mean retrieval score, mean *strength*); Solid line:$\gamma = \gamma_1$, Dotted line: $\gamma = \gamma_2$ where $\gamma_1 < \gamma_2$; Mean of each curve given in legends.

Notice that this evaluation measure takes into account the *type* assigned to each identified relation based on a word group describing a context in the news articles.

6.3 Automatic Evaluation Results

In this section, we discuss the results of automatic evaluation of our system. We start by discussing the impact of the parameter γ of the system. As discussed in Section 3.2.3, increasing the value of γ forces the system to pick named entities that have stronger evidence from word groups.

Figures 6 and 8 show the effect of γ on the evaluation measures for the TIME and BBC datasets, respectively. In these figures, the x-axes represent the indices of one-month–long time intervals (TIME dataset) or indices of random subsets of the data (BBC dataset). One semantic network is built for each time interval or subset. The y-axes in these figures give the magnitude of various evaluation measures. The dotted lines are for a higher value of γ as compared to the solid lines. It is observed that the solid line is higher

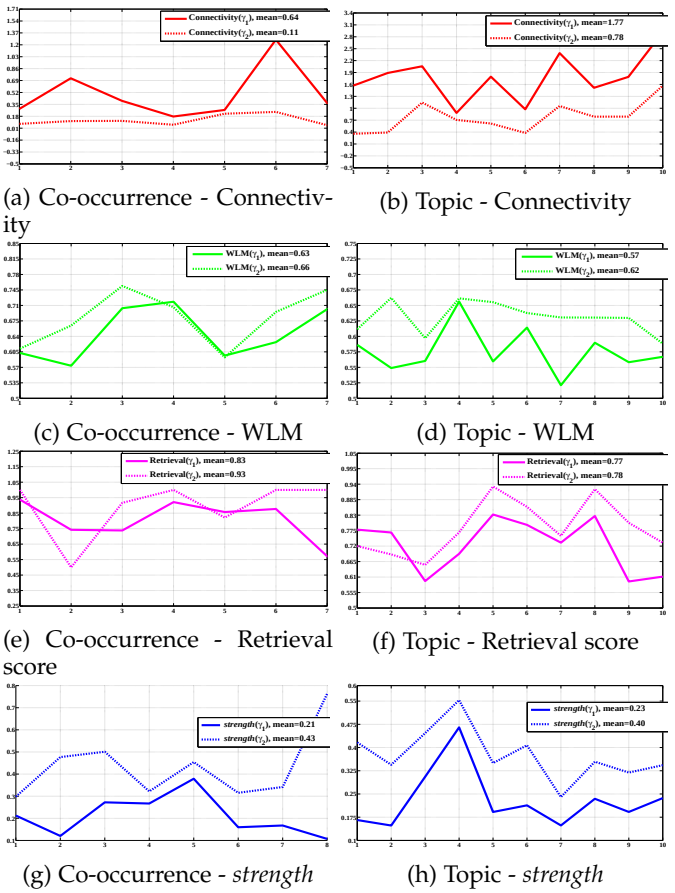


(a) Connectivity (b) WLM

(c) Retrieval score (d) *Strength*

Fig. 7: Summary statistics (boxplots) for connectivity, mean WLM, mean retrieval score, and mean *strength* at γ_2 minus that at γ_1 over all time periods(TIME dataset), $\gamma_2 > \gamma_1$

than the dotted line for mean connectivity, as increase in γ produces fewer relations. The dotted line is generally higher than the corresponding solid line for mean WLM, mean retrieval score, and mean *strength*, as increase in γ



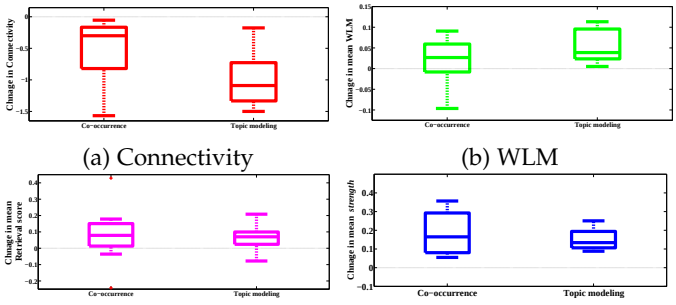
(a) Co-occurrence - Connectivity (b) Topic - Connectivity

(c) Co-occurrence - WLM (d) Topic - WLM

(e) Co-occurrence - Retrieval score (f) Topic - Retrieval score

(g) Co-occurrence - *strength* (h) Topic - *strength*

Fig. 8: Effect of threshold γ on evaluation measures on BBC dataset; x-axis: Dataset sample, y-axis: Evaluation measure (connectivity, mean WLM, mean retrieval score, or mean *strength*); Solid line:$\gamma = \gamma_1$, Dotted line: $\gamma = \gamma_2$ where $\gamma_1 < \gamma_2$; Mean of each curve given in legends.



(a) Connectivity (b) WLM

(c) Retrieval score (d) *Strength*

Fig. 9: Summary statistics (boxplots) for connectivity, mean WLM, mean retrieval score, and mean *strength* at γ_2 minus that at γ_1 over all samples (BBC dataset), where $\gamma_2 > \gamma_1$

forces the system to pick relations with stronger evidence. These trends are consistent across both datasets and all configurations of the system for every type of word group (co-occurrence, keyword, topic). Figures 7 and 9 depict the summary statistics for change (with increase in γ) in mean evaluation measures of semantic networks built for all time intervals for the TIME dataset and all subsets of the BBC dataset, respectively. Each boxplot shows the minimum, 25th percentile, 50th percentile (median), 75th percentile, and maximum of the change in the corresponding mean evaluation measure. A positive value indicates an increase in the mean evaluation measure. It is clear from these figures that increase in mean WLM, retrieval score, and *strength* with increase in γ is the dominant trend as corresponding boxplots are above the zero-line.

Note that the trend of change in mean evaluation measures for each semantic network is stronger in the TIME dataset (Figure 6) than that in the BBC dataset (Figrue 8). This can be attributed to the fact that semantic networks on the TIME dataset are generated on articles published during one month, thus ensuring that many of the articles of one news story are available for processing together and resulting in more meaningful relations between named entities in the context of that story. On the other hand, the networks on the BBC dataset are generated on random subsets of the data with no guaranty of availability of significant information about one news story in one subset.

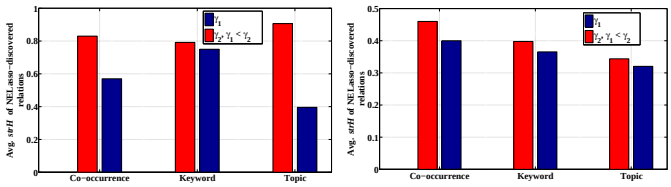
The parameters τ and K (discussed in Sections 4.2.1, 4.2.2 and 4.2.3) control the number of word groups. We observed that fewer word groups of larger sizes generate more relations increasing connectivity of the semantic networks. However, the threshold γ affects both the number and the quality of the discovered relations regardless of the word groups' size and count. Thus, the choice of γ is more important than that of τ and K in our system.

Comparing different word group formation methods, we observe that topic-based word groups tend to generate higher numbers of relations than co-occurrence- and keyword-based word groups, for the same value of γ. This can be due to the fact that only topic-based word groups are overlapping, thus making more named entities to share word groups with higher positive evidence. The mean *strength* assigned to the discovered relation is generally higher for topic-based word groups than all other group formation methods for comparable values of connectivity and quality measures (WLM and retrieval score). This is because threshold γ is set to a higher value for topic-based word groups to generate about the same number of relations as other methods.

It is necessary for generated word groups to correspond to relevant contexts or specific stories in news material, rather than to syntactic categories, for the discovered relations to make sense to news readers. We experimented with Brown clustering method that clusters words according to their syntactic behavior [34]. However, the resulting named entities relations did not fare well in our evaluations.

6.4 Human Evaluation Study

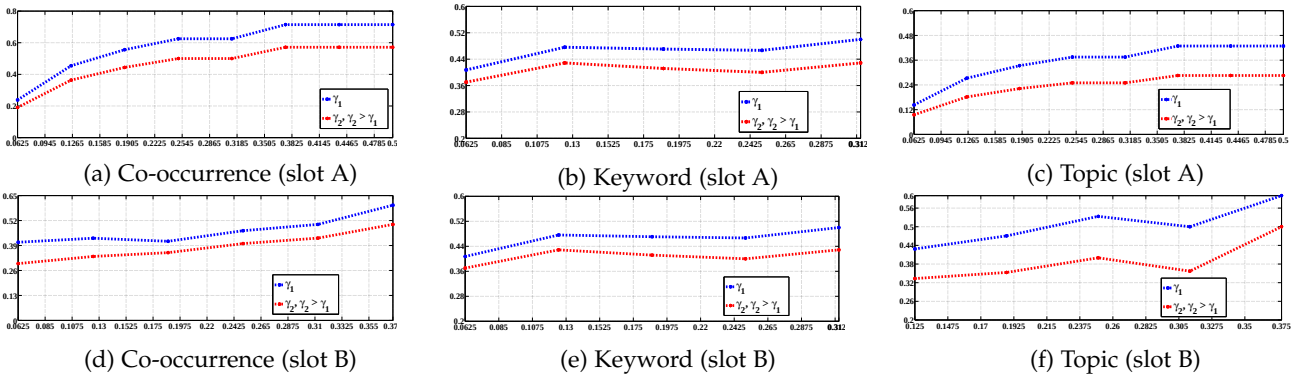
We also conducted a human evaluation study of our system. The aim of this study is to compare the semantic network



(a) Slot A (b) Slot B

Fig. 10: Human evaluation of NELasso; height of each bar represents mean of human-assigned strength to relations discovered by NELasso; Blue:$\gamma = \gamma_1$, Red: $\gamma = \gamma_2$ such that $\gamma_1 < \gamma_2$

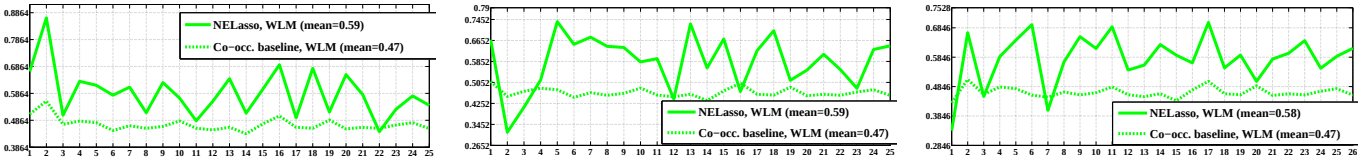
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2632117, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Co-occurrence (slot A) (b) Keyword (slot A) (c) Topic (slot A)

(d) Co-occurrence (slot B) (e) Keyword (slot B) (f) Topic (slot B)

Fig. 11: Fraction of human-identified relations discovered by NELasso; x-axis: Minimum $strH(r_{ij})$ of human-identified relations, y-axis: fraction of human-identified relations discovered by NELasso; Blue:$\gamma = \gamma_1$, Red: $\gamma = \gamma_2$ such that $\gamma_1 < \gamma_2$



(a) Baseline vs. Co-occurrence-based NELasso (b) Baseline vs. Keyword-based NELasso (c) Baseline vs. Topic-based NELasso

Fig. 12: Comparison between co-occurrence-based baseline model and various configurations of NELasso using WLM (TIME dataset); x-axis: Time period (month), y-axis: mean WLM; Mean of each curve given in legends.

built by our system against that built by humans when given the same set of news articles. This study is conducted on the TIME dataset.

In our human evaluation study, we fixed the duration of time slot to one day so as to limit the number of articles to a number easily readable by human judges. We selected two time slots, referred to as slot A and slot B, containing 10 and 8 articles, respectively. We presented the judges with a matrix of named entities for each time slot, such that each cell of the matrix corresponds to the pair of entities indicated by the row and the column. The judges were asked to read the articles for each time slot and mark in the matrix whether or not each pair of named entities is related based on the articles in the time slot. We collected observations from 16 judges. Likewise, NELasso was employed to automatically build semantic networks of named entities for slot A and slot B.

The strength of a relation in human evaluation is estimated from the number of judges who mark that relation. When a relation is identified by many judges, it indicates that the relation is clear and strong enough to be recognized readily by humans. Accordingly, the strength, $strH_{ij}$, of a relation between entities e_i and e_j is defined as

$$strH(r_{ij}) = \frac{\text{No. of judges that identify } r_{ij}}{\text{Total no. of judges}} \quad (7)$$

Figure 10 shows that the $str(r_{ij})$ assigned to a relation between entities e_i and e_j by our system is generally a good indicator of $strH(r_{ij})$, i.e., the strength assigned to the relation by humans. As threshold γ is increased, NELasso identifies fewer but stronger relations. It is seen from Figure 10 that the fewer relations at higher γ also have higher mean $strH(r_{ij})$ than those selected by the lower γ value. The strength assigned to relations by NELasso correlates well with that assigned by humans.

It is also important to check how many of the relations identified by the judges are discovered by NELasso. Our system is able to discover higher fractions of human-identified relations with higher $strH(r_{ij})$ (Figure 11). The blue and red lines indicate lower and higher values of threshold γ, respectively. The horizontal axis shows the minimum $strH(r_{ij})$ of relations identified by humans. In general, with increasing minimum $strH(r_{ij})$, larger fractions of human-identified relations are discovered by NELasso (given on y-axis). This trend is more pronounced in slot A than in slot B.

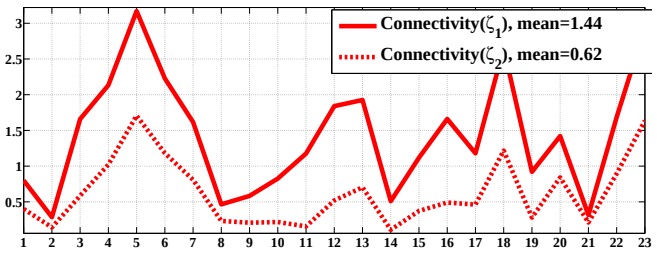
We also compute the Fleiss-kappa for human judges, which is an effective measure for inter-rater reliability [33]. Fleiss-kappa for slot A is 0.6 which reflects moderate-to-substantial inter-rater agreement. Fleiss-kappa for slot B is 0.37 which indicates fair agreement among the judges.

6.5 Co-occurrence-based Baseline Model

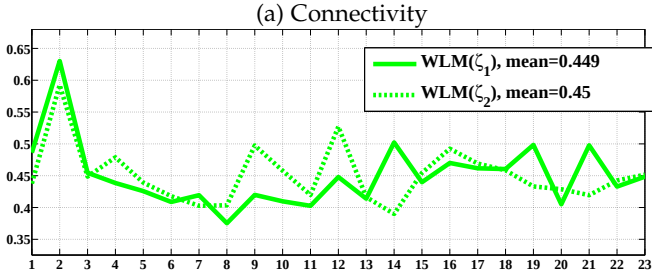
In this section, we present a baseline model for finding relations between named entities. According to this model, a relation exists between two named entities when they co-occur in the same article. We compare the quality of relations found by NELasso and by this baseline model using WLM on different time periods of the TIME dataset.

Figure 12 shows that mean WLM for relations found in each month by any configuration of NELasso is much higher than that of the baseline model. This confirms that the straightforward method of constructing relations based on co-occurrence of entities in articles generates a large quantity of substandard relations with no information regarding their type or statistical signature.

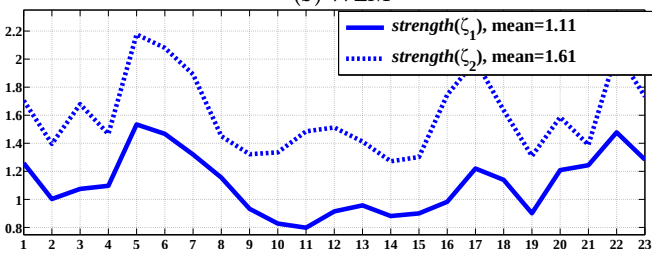
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2016.2632117, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Connectivity



(b) WLM



(c) *Strength*

Fig. 13: Effect of threshold ζ of linear model baseline system on evaluation measures (TIME dataset); x-axis: Time period (month), y-axis: Evaluation measure (connectivity, mean WLM, mean retrieval score, mean *strength*); Solid line:$\zeta = \zeta_1$, Dotted line: $\zeta = \zeta_2$ where $\zeta_1 < \zeta_2$; Mean of each curve given in legends.

6.6 Value of Sparse Group Learning

In this section, we address a fundamental question regarding our model: what is the benefit of sparse group learning over standard un-regularized learning?

To answer this question, we consider another baseline system that learns simple linear models for all named entities by using their positive and negative examples (i.e., articles) in equal numbers. This system learns a coefficient vector $\mathbf{p}_i \in \mathcal{R}^N$ for the ith named entity, e_i. The relation $rel(e_i, e_j)$ between entities e_i and e_j is decided based upon the cosine similarity between $\mathbf{p}_i$ and $\mathbf{p}_j$; if this similarity is greater than a threshold ζ, the system declares that there is a relation r_{ij} between entities e_i and e_j with *strength* $str(r_{ij}) = \mathbf{p}_i^{\mathrm{T}}\mathbf{p}_j$. There is no way of finding a meaningful *type* for this relation.

Figure 13 shows mean connectivity, WLM, and *strength* for semantic networks built by the linear model baseline system for each time interval of TIME dataset. Note that there is negligible change in mean WLM even after significant change in mean connectivity of the network. The difference between mean WLM as we change ζ is also negligible while the corresponding difference in NELasso is substantial (Figure 6). For time intervals where mean WLM changes with increase in threshold ζ, often the change is negative indicating deterioration in the quality of discovered relations. This implies that there is little correlation between relation *strength* in the linear model baseline system and the quality of the identified relations, as judged by WLM. This is the first advantage of employing group sparse learning in the proposed system.

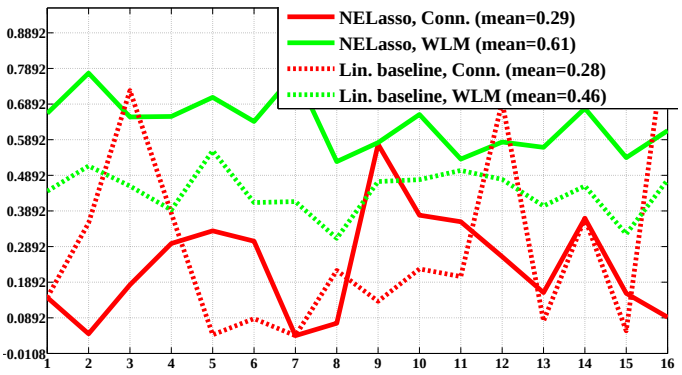


Fig. 14: Comparison between NELasso and linear model baseline system (TIME dataset); x-axis: Time period (month), y-axis: Evaluation measure (connectivity, mean WLM); Mean of each curve given in legends.

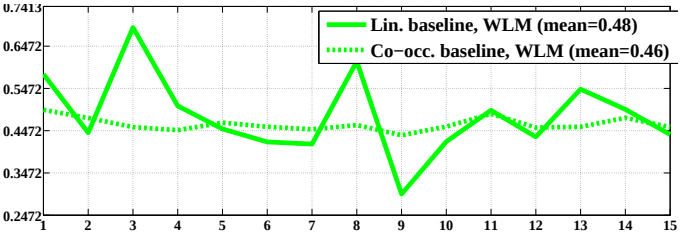


Fig. 15: Comparison between co-occurrence and linear model based baselines on the basis of WLM (TIME dataset); x-axis: Time period (month), y-axis: mean WLM for relations; Mean of each curve in legend.

Figure 14 highlights another advantage of NELasso over the simple linear model based baseline system. When the two systems find almost similar numbers of relations among named entities, the relations identified by sparse group learning are of much higher quality than those identified by the linear model based system. Furthermore, the sparse group learning based system assigns a meaningful relation *type* to each identified relation. No meaningful relation *type* can be identified in simple linear modeling based baseline.

The simple linear model performs only slightly better than the co-occurrence-based model presented in Section 6.5. Figure 15 shows that mean WLM of relations found by the linear model is higher than that of co-occurrence-based model for a few time intervals only. In comparison, NELasso performs consistently better than both baseline models in terms of mean WLM as shown in Figure 12.

6.7 Sensitivity Analysis

Our system requires that a few parameters are set before its execution. The parameters include the weights λ_1 and λ_2 assigned to the two penalty terms involved in the group sparse logistic regression model and the threshold γ on relation *strength* for its selection. We study experimentally

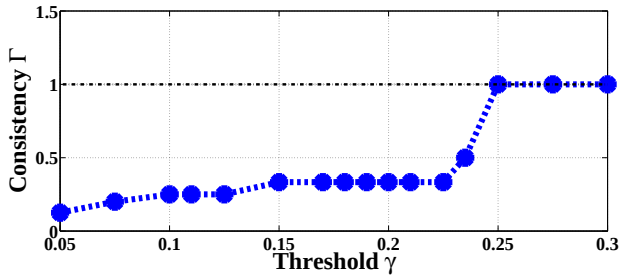


Fig. 16: Effect of threshold γ on the consistency of system's output (TIME dataset)

the effect of these parameters on the system's output. Furthermore, we also study the impact of sampling of negative articles on system output.

The parameter γ controls the selection of relations such that only relations with $str_k(r_{ij}) = t_k^{e_i} \times t_k^{e_j}$ with $t_k^{e_i} > \gamma$ and $t_k^{e_j} > \gamma$ are included in the output (refer to Definition 1 for details). As γ is increased, fewer relations of higher *strength* (quality) are selected for inclusion in the semantic network. Moreover, relations with high *strength* are unaffected by significant increase in γ. This trend is discussed in detail in Section 6.3 (see Figures 6 and 8).

We observe that the increase in values of λ_1 and λ_2 has the same effect as increase in the value of γ. As λ_1 and λ_2 are increased, more emphasis is put on sparsity of the logistic regression model, i.e., entries of coefficient vector $\mathbf{x}$ become smaller and a larger number of them are set to 0. Since the relations between named entities are decided based on sum of entries x_n of $\mathbf{x}$ such that nth word belongs to a certain word group (see Section 3.2.3), fewer relations are discovered with the increase in values of these parameters. But, relations with high *strength* are unaffected as the sum of coefficients of the named entities wil be higher than those for other entities. Of course, the threshold γ has to be adjusted downward since the absolute strength value will be lower. Due to lack of space, we do not show the variation curves for λ_1 and λ_2.

While learning the group sparse logistic regression model for a named entity, our system randomly selects a set of articles that do not mention the named entity (since the number of articles that mention an entity is much smaller than those that do not). We study the sensitivity of the system's output to selection of different set of negative examples by evaluating the output from five runs of the system. Each time, the system randomly selects sets of negative examples for each entity. Let X_t be the set of relations discovered in the tth iteration out of a total of T iterations, then the *consistency* of the systems output is defined as

$$consistency = \frac{|\cap_{t=1}^{T} X_t|}{|\cup_{t=1}^{T} X_t|} \tag{8}$$

In words, consistency is the ratio of the number of common relations found in all iteration to the number of unique relations found in all iterations. Its maximum value is 1, i.e., all relations are discovered in all iterations.

Figure 16 shows the effect of γ on system consistency. It is observed that as γ is increased the systems output becomes more and more consistent until its consistency reaches 1. This implies that relations of high *strength* are consistently discovered in all iterations despite variations in selection of negative examples.

6.8 Time Complexity and Scalability

NELasso is not only effective but also time efficient and scalable to large-scale applications of identifying relations among named entity from published news articles automatically. The system identifies word groups once for a set of articles and uses them while learning a sparse logistic regression model for each named entity mentioned in that set of articles. Our system takes on average 0.05 seconds to process one named entity on a machine with 3.40 GHz processor and 32 GB memory. It is clear that our system can scale up easily to practical settings involving large sets of news articles collected from multiple sources on daily basis.

7 Conclusion

Named entities are an important aspect of online textual content such as news articles. While they are subjects of interest for many on the Web, relations between named entities provide additional insights into evolving trends in news stories. Reliable automatic discovery of relations between named entities from large collections of data can help improve news recommender systems and search-navigation experience of users.

In this paper, we present a novel system for understanding named entities in their contexts through a sparse structured model. We exploit additional knowledge of contexts, such as keywords and topics, to define a group structure over the words, which in turn enables us to specify both the type and the strength of the discovered relations between named entities. Our system is unsupervised and requires minimal parameter settings, and outputs an informative network of entities mentioned in the processed collection. Unlike many previously proposed frameworks that aim at producing a database of facts, our system aims at machine understanding of news collections without requiring external resources or initial seeds of named entities. It also imposes no restriction on the types of relations to be identified.

Our experiments on two real-world news collections demonstrate the ease with which relations between named entities can be tracked over time. We also conduct an extensive evaluation of our system, including human assessment and comparisons with two baseline systems. The results show that our system performs consistently well across various configurations.

This work demonstrates the power of sparse structured learning in an entirely new setting. With careful modeling, this approach can also be applied to other problems to yield effective unsupervised solutions. There is much potential in incorporating other kinds of knowledge like authorship or polarity within a structured learning framework for enhanced understanding of textual content.

References

[1] G. Mishne and M. De Rijke, "A study of blog search," in *Advances in information retrieval*, 2006, pp. 289–301.

[2] E. Gabrilovich, S. Dumais, and E. Horvitz, "Newsjunkie: providing personalized newsfeeds via analysis of information novelty," in *Proceedings of the 13th international conference on World Wide Web*.

[3] S. Szumlanski and F. Gomez, "Automatically acquiring a semantic network of related concepts," in *Proceedings of the 19th ACM international conference on Information and knowledge management*.

[4] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on Natural language processing of the AFNLP*.

[5] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *Proceedings of the fifth ACM conference on Digital libraries*.

[6] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. Strassel, and R. M. Weischedel, "The automatic content extraction (ace) program-tasks, data, and evaluation," in *LREC*, 2004.

[7] T. Hirano, Y. Matsuo, and G. Kikui, "Detecting semantic relations between named entities in text using contextual features," in *Proceedings of the 45th annual meeting of the ACL on Interactive poster and demonstration sessions*.

[8] D. T. Bollegala, Y. Matsuo, and M. Ishizuka, "Relational duality: Unsupervised extraction of semantic relations between entities on the web," in *Proceedings of the 19th international conference on World wide web*.

[9] J. Zhu, Z. Nie, X. Liu, B. Zhang, and J.-R. Wen, "Statsnowball: a statistical approach to extracting entity relationships," in *Proceedings of the 18th international conference on World wide web*.

[10] F. Wu and D. S. Weld, "Open information extraction using wikipedia," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

[11] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[12] M. Schmitz, R. Bart, S. Soderland, O. Etzioni *et al.*, "Open language learning for information extraction," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[14] L. Meier, S. Van De Geer, and P. Bhlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[15] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[16] J. Yin, X. Chen, and E. P. Xing, "Group sparse additive models," in *Proceedings of the 29th international conference on Machine learning*, 2012, pp. 871–878.

[17] X. Chen, X. Shi, X. Xu, Z. Wang, R. Mills, C. Lee, and J. Xu, "A two-graph guided multi-task lasso approach for eqtl mapping," in *International conference on Artificial intelligence and statistics*, 2012, pp. 208–217.

[18] D. Yogatama, Y. Sim, and N. A. Smith, "A probabilistic model for canonicalizing named entity mentions," in *Proceedings of the 50th annual meeting of the Association for computational linguistics*.

[19] C. Corley and R. Mihalcea, "Measuring the semantic similarity of texts," in *Proceedings of the ACL workshop on Empirical modeling of semantic equivalence and entailment*.

[20] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871–882, 2003.

[21] A. F. Martins, N. A. Smith, P. M. Aguiar, and M. A. Figueiredo, "Structured sparsity in structured prediction," in *Proceedings of the conference on Empirical methods in natural language processing*.

[22] Y. Jin, E. Kiciman, K. Wang, and R. Loynd, "Entity linking at the tail: sparse signals, unknown entities, and phrase models," in *Proceedings of the 7th ACM international conference on Web search and data mining*.

[23] Y. Jin, K. Wang, and E. Kiciman, "Sparse lexical representation for semantic entity resolution," in *IEEE international conference on Acoustics, speech and signal processing (ICASSP), 2013*.

[24] W. Shen, J. Wang, P. Luo, and M. Wang, "Linden: linking named entities with knowledge base via semantic knowledge," in *Proceedings of the 21st international conference on World Wide Web*.

[25] R. Navigli and S. P. Ponzetto, "Babelnet: Building a very large multilingual semantic network," in *Proceedings of the 48th annual meeting of the Association for computational linguistics*, 2010, pp. 216–225.

[26] X. Dai, J. Jia, L. El Ghaoui, and B. Yu, "Sba-term: Sparse bilingual association for terms," in *Fifth IEEE international conference on Semantic computing, 2011*.

[27] B. Rosenfeld and R. Feldman, "Clustering for unsupervised relation identification," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 411–418.

[28] T. Hasegawa, S. Sekine, and R. Grishman, "Discovering relations among named entities from large corpora," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 415.

[29] A. Tariq and A. Karim, "Fast supervised feature extraction by term discrimination information pooling," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 2233–2236.

[30] K. N. Junejo and A. Karim, "A robust discriminative term weighting based linear discriminant method for text classification," in *IEEE International Conference on Data Mining (ICDM)*, 2008, pp. 323–332.

[31] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[32] I. Witten and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in *Proceeding of AAAI workshop on Wikipedia and artificial intelligence: an Evolving synergy*, 2008, pp. 25–30.

[33] L. Fleiss, B. Levin, and M. C. Paik, "The measurement of interrater agreement," in *In Statistical methods for rates and proportions (2nd ed*, 1981.

[34] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J.D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," in *Cmputational Linguistics 18, no. 4*, 1992.

Amara Tariq is a Ph.D. student at Computer Science division of College of Electrical Engineering and Computer Science at University of Central Florida (UCF). Her research interests include machine learning, image processing and natural language understanding. She was awarded Fulbright fellowship for doctoral studies in 2011.

Asim Karim (M'97) received his B.Sc. (Honors) Engineering degree from University of Engineering and Technology (UET) Lahore, Pakistan, in 1994 and his M.S. and Ph.D. degrees from The Ohio State University in 1996 and 2002, respectively. He is currently Professor of Computer Science at Lahore University of Management Sciences (LUMS) where he directs the Knowledge and Data Engineering research group. His research interests include data mining and machine learning with recent focus on text analytics. He is the author/co-author of two books and over 50 research publications. His works have won recognition at the 2006 and 2007 ECML/PKDD Discovery Challenge, and he is the recipient of the 2015 PAS-COMSTECH Prize in Computer Science from the Pakistan Academy of Sciences.

Hassan Foroosh (M02 - SM03) is a Professor in the Department of Electrical Engineering and Computer Science at the University of Central Florida (UCF). He has authored and co-authored over 100 peer-reviewed journal and conference papers, and has been in the organizing and the technical committees of many international conferences. Dr. Foroosh is a senior member of IEEE, and an Associate Editor of the IEEE Transactions on Image Processing since 2011. He was also an Associate Editor of the IEEE Transactions on Image Processing in 2003-2008. In 2004, he was a recipient of the Pierro Zamperoni award from the International Association for Pattern Recognition (IAPR). He also received the Best Scientific Paper Award in the International Conference on Pattern Recognition of IAPR in 2008. His research has been sponsored by NASA, NSF, DIA, Navy, ONR, and industry.