Dealing with biographical information (e.g., biography generation, answering biography-related questions, etc.) requires the identification of important activities in the life of the individual in question. While there are activities that can be used in any biography (e.g., person was born on a particular date, person lived in a particular location, etc.), many activities used in biographies tend to be occupation-related, others are person-specific. Hence, occupation gives important clues as to which activities should be included in the biography. In this paper, we present a methodology for identifying a three-level hierarchy of biographical activities: those activities that apply to the general population, those activities that are occupation-related, and those activities that are person-specific. We use the obtained occupation-related activities as features for a multi-class SVM classifier to identify the occupation of a previously unseen individual. We also show that the activities automatically obtained from text can be used as features not only for a classification task but for a clustering task as well. We show that, given the correct number of clusters, people belonging to the same occupation are clustered together. At the same time, clustering people into a smaller number of classes allows the grouping of practitioners of the occupations that share a considerable number of occupation-related activities. Thus, analyzing descriptions of people belonging to various occupations, we can build a hierarchy of occupations.

© 2012 Elsevier B.V. All rights reserved.

Keywords: Biography information Occupation classification

1. Introduction

Natural language processing (NLP) information selection applications such as summarization and question-answering (QA) systems are designed to reduce the amount of time necessary for finding information of interest. Summarization systems produce a condensed version of the generally important information presented in the input, while QA systems target specific information according to a certain question.

Recently there has been increased interest in creating systems which combine summarization and QA. Knowing the domain for which the answer should be constructed is crucial as it guides the information selection process. For example, the understanding that the output text should be about an earthquake, or a definition of a certain object, or an opinion piece, or a biography, etc. can be used for designing a better information selection approach that outperforms general summarization techniques. Such domain specific approaches take advantage of the fact that the selected information might not be important in general but targets specific piece of information that is crucial for the domain under analysis.

In our work we concentrate on the biography domain which is a popular domain to work with not only in the field of NLP but also in the related field of Information Retrieval (IR). In this work, we describe a methodology for identifying the text constructions that...
are used for describing human activities. Moreover, we hypothesize that human activities can be divided into three categories: biographical facts (the person’s place and date of birth, where the person lived, etc.); activities typically associated with the person’s occupation (e.g., singers sing, explorers travel around the globe to study new lands, artists create paintings); and person-specific activities that make this person unique and distinguish her from other people and/or practitioners of the same occupation. We describe a methodology for grouping the extracted activities into these three groups. The performance of this methodology is evaluated through the classification task where we use the identified occupation-related activities to predict the occupation of a person. Existing research shows that knowing the person’s occupation is helpful for detecting information which should be used in the biography [18,47].

Predicting a person’s occupation can be formulated as a regular classification task. However, our way of extracting classification features and placing these features into a hierarchy is novel. We also evaluate the quality of the created hierarchy through a clustering task: we use the extracted activities to group people according to their occupations. We empirically show the value of our hierarchical approach by grouping people performing similar activities and thus, creating a hierarchy that can be used by information selection system dealing with the biography domain.

We believe that the presented methodology for identifying text constructions that are descriptive for a particular domain and placing these constructions into a domain hierarchy goes far beyond dealing with the biography domain studied in this paper. Rather, this methodology can be applied to automatic generation of domain hierarchies. These hierarchies, in their turn, can be used by information selection systems to improve the quality of their output.

1.1. Our contributions

To summarize, the contributions in this paper are as follows:

- We describe a novel approach for identifying in text human biographical activities and for dividing these activities into three levels: general biographical, occupation-related, and person-specific activities. This three-level division is performed using groupings of documents as well as the HITS-based ranking Kleinberg [33].
- We use the extracted activities as classification features to predict an occupation of a person. The predicted occupation can be used as a guideline for a variety of information selection tasks (e.g., summarization, QA, IR). We first described this classifier in Filatova and Prager [23].
- We also use the extracted activities as clustering features to group those people who perform similar activities and thus are likely to belong to the similar, or related, or same occupation. As for any clustering task, the choice of the number of output clusters is crucial. In this paper, we do not attempt to identify the correct number of clusters. However, we show that, given the correct number of clusters, people belonging to the same occupation are clustered together. Moreover, clustering people into a smaller number of clusters allows grouping of practitioners of those occupations that share a considerable number of occupation-related activities.

We believe that the contributions of our work are not limited to the biography domain. Rather, the described approaches can be used for learning interesting facts about many domains.

1.2. Organization of the paper

The rest of the paper is organized as follows. In Section 2, we describe the related work. In Section 3, we describe how we create a balanced corpus containing practitioners of ten occupations. We then describe the procedure we designed and implemented for learning biographical activities used in descriptions of people (Section 4). In Section 5, we show how random walk theory can be used to rank extracted activities in the order of their importance for a particular occupation. In particular, we use the Hubs and Authorities (HITS) algorithm suggested by Kleinberg [33] for ranking web pages. We then demonstrate how this ranking can be used to divide the activities extracted from descriptions of people belonging to the same occupation into three classes: general biographical, occupation-related and person-specific activities. In Section 6, we use the identified occupation-related activities as classification features and evaluate a classifier that assigns a person to a particular occupation based on the occupation-related activities used in the description of this person. We run this classification experiment using two different corpora and several baselines that are widely used in the field of NLP. In Section 7, we discuss the results of the classification experiments for the both corpora. In Section 8, we show that the activities extracted from people’s descriptions can be used not only for classification but for clustering as well. Finally, in Section 9, we conclude our discussion and describe the avenues for future research.

2. Related work

Recently there has been increased interest in creating systems that output answers to open-ended questions (including, biography generation systems) which combine summarization and QA and can give long answers to definition, biography, opinion and other question types. Such systems employ general summarization techniques and at the same time, take advantage of the fact that the output does not target generally important information but rather covers certain types of facts related to the question type.
For example, the systems participating in DUC 2004 created summaries which could be used as answers to the question “Who is X?” These systems use a wide variety of techniques. Blair-Goldensohn et al. [11] treat “Who is X?” as a definition question and use the DeFscriber system [12] to create focused summaries that would correspond to biographies. The DeFscriber system uses a combination of goal-driven and data-driven techniques. First, definition predicates are used to select information suited for a definition (e.g., genus-species information), and the rest of the answer is shaped according to the themes found in the input document collections.

Biryukov et al. [10] use Topic Signatures of Lin and Hovy [38] constructed around the person’s name, as Topic Signature generation exploits the natural tendency of the semantically related words to co-occur more often than by chance in the same context.

Zhou et al. [55] present a system created specifically for the biography generation task. It uses nine features which are likely to be used in biography texts: bio (biographical facts), fame, personality, social, education, nationality, scandal, personal, work. Using manually annotated 130 biographies they learn the textual patterns corresponding to these nine features. We suggest an algorithm that can be used for unsupervised learning of the features similar to the nine manually defined biography-related features used in Zhou et al. [55].

Extraction of biographical activities has been the focus of many Information Extraction (IE) systems. Recently, NIST has created a new evaluation effort for IE event identification, Automatic Content Extraction (ACE) [17]. For the ACE task, the participating systems are supposed to identify several pre-defined semantic types of events (life, justice, transaction, conflict, etc.) together with the constituent parts corresponding to these events (agent, object, source, target, time, location, other). Among the ACE pre-defined events are life events. Table 1 lists life events together with the arguments which should be extracted for these events.

Biadsy et al. [9] propose to use the uniform structure of biographical Wikipedia articles to learn important biographical activities that should be included in the answer to the “Who is X?” question.

Garera and Yarowsky [24] suggest several approaches for learning biographical activities from text. One of the suggested approaches is to use transitive models over attributes via consensus from attributes of neighboring names. “Attributes such as Occupation are transitive in nature, that is, the people names appearing close to the target name will tend to have the same occupation as the target name.”

Biographical information can be used to answer questions other than the “Who is X?” type. Prager et al. [44] use biographical information within their QA-by-Dossier-with-Constraints system, which checks whether the possible answer satisfies the constraints for the person about whom the question is asked. One of such natural constraints for artists, composers and writers is that all their works are produced in the span of time between the dates of birth and death. For example, for the question “When did Leonardo da Vinci paint the Mona Lisa?” five candidate answers with their initial confidence scores are presented in Table 2.2

It must be noted that the initial scores are computed without taking into consideration the biographical information. The correct answer 1503 is scored fourth. However, using Leonardo da Vinci’s biographical information, namely that he was born in 1452 and died in 1519, makes only two answers from Table 2 possible. Choosing out of these two possible answers the one with the highest score gets the correct answer to the questions “When did Leonardo da Vinci paint the Mona Lisa?”

Knowing that biographical activities can be used within information retrieval (IR) applications as a significant percentage of search queries seem to contain references to people. Guha et al. [27] describe search over a semantic web where “the semantic web is not a web of documents, but a web of relations between resources denoting real world objects.” They show an advantage of having real world objects linked to each other. For example, when asked about ⟨Yo-Yo Ma⟩ semantic search outputs a network of objects. This network encodes information that Yo-Yo Ma is a [musician], he is connected to ⟨Paris, France⟩ by the link ⟨birth place⟩, to ⟨10/07/55⟩ by the link ⟨birth date⟩, etc. Fig. 1 shows a small chunk of the Semantic Web corresponding to the cellist Yo-Yo Ma.3

Another IR task that can benefit from biographical information extraction is the People Search task [4–6]. The goal of the People Search task is, given the results for a person name search that contain a mix of documents about different people sharing the same name, divide these documents into clusters each of which describes one person.

Learning features important for descriptions of entities of a particular type can be applied not only to biography generation but to dealing with other domains as well. For example, Filatova et al. [22] suggest a feature selection approach using document collections describing various domains (i.e., earthquakes, presidential elections, etc.) to learn domain templates that can be used

---

2 Table 2 is from Prager et al. [44].
3 Fig. 1 is from Guha et al. [27].
for a variety of tasks, including domain-dependent summarization. Sauper and Barzilay [46] propose an algorithm for template induction from human-authored documents; these templates are used for automatic generation of Wikipedia articles.

3. Data

For the experiments described in this paper we create our own set of people belonging to various occupations as we are aware of no set diverse enough to analyze activities of people belonging to different occupations. We decided not to use the list of people whose biographies were created for DUC 2004 task as the input documents for this task are contemporary newswire articles and more than a half of the 50 people used there were politicians.

We created our own corpus of biographies. First, we decided on 10 occupations. We deliberately chose a diverse set of occupations, so that it includes occupations that are similar in nature (e.g., mathematicians and physicists) and occupations whose practitioners are not likely to have overlapping activities (e.g., mathematicians and dancers). We understand that 10 occupations do not cover all possible occupations, but that is not critical to our study.

Our next step was to decide on 20 practitioners for each occupation. As described later, we sought documents for each of the practitioners for every occupation. Since no documents were found for some of the individuals, these people were eliminated from the experiments; 189 survived.

To extract occupation-related activities (see Section 4) we created 189 text collections, one per person of interest. Documents in each of these collections were fetched by search and either describe the person or at least mention his/her name. Each person belongs to one of the ten pre-defined occupations. We also had to make sure that the people included into the final lists were mentioned in a sufficient amount of documents in our corpora.

The document collections we created covered:

- 20 artists;
- 18 athletes;
- 20 composers;
- 15 dancers;
- 17 explorers;
- 20 mathematicians;
- 19 physicists;
- 20 politicians;
- 20 singers;
- 20 writers.

We observed that for some occupations humans generally agree upon its representative labels, despite the fact that the activities performed by these representatives might be very different. For example, no matter to what school an artist belongs (impressionism, surrealism) he/she is usually addressed as an artist. The situation with politicians is different. They are often referred to not as politicians, but according to their political parties (e.g., democrat, republican, socialist) or by the post held (e.g., president, prime-minister). Choosing an appropriate occupation title becomes crucial at the document retrieval stage as this title is used in the search engine query.

Our occupation list satisfies the following criteria that are important for our study:

- it is diverse and covers a substantial variety of occupations, including arts, sciences and other aspects of human activities
- it contains some occupations which are closely related; for example, mathematicians and physicists
- it contains occupations that are very different and in which it is almost impossible to specify activities which are routinely performed in both; for example, singers and explorers.

To get the lists of people belonging to each particular occupation we use WordNet 2.0 Fellbaum [20] (e.g., hyponyms for composer contain a list of composers). We also use “Google Sets” interface, 4 which has been successfully used to find people belonging to the same occupation [44]. It is possible to use bootstrapping to get more practitioners for every occupation. However, automatic addition of people identified by bootstrapping (or any other automated approach) would have necessarily contained

---

4 http://labs.google.com/sets.
noise, which would have made the reported results not accurate. As both the task and the approaches described in this paper are novel, it is important for us to keep the results as clean and accurate as possible. Also, we believe, that given more data points, the approach type that we use in this paper will only show higher performance.

It must be noted, that though we work with a balanced corpus where every occupation has approximately the same number of practitioners, the real-life distribution of people among occupations is not balanced. However, the task of predicting or estimating the real-life distribution of practitioners among occupations is very difficult and goes beyond the scope of this paper. Moreover, according to Weiss and Provost [51]: “if no additional information is provided and a class distribution must be chosen without any experimentation, our results show that for accuracy and for AUC maximization, the natural distribution and a balanced distribution (respectively) are reasonable default training distributions.”

For the first set of experiments we retrieve documents from four corpora: AQUAINT, TREC, part of World Book and part of Encyclopedia Britannica. For document retrieval we use IBM’s JuruXML search engine [14] which indexes tags such as named entity class labels along with the text itself. Queries to JuruXML may include tagged terms; these will only match similarly tagged instances in the index. We make use of named-entity labels

\[\text{\langle person\rangle}\text{Last_Name}/\text{\langle person\rangle}\text{\langle role\rangle}\text{Occupation}/\text{\langle role\rangle}\text{.}\]

likewise in the queries to perform two types of word sense disambiguation: (1) to differentiate a person from, for example, a location with the same name (e.g., Newton – a physicist and Newton – a town in Massachusetts); and (2) to differentiate two different people with the same name belonging to different occupations (e.g., Louis Armstrong a singer and Lance Armstrong an athlete). These issues can be partially avoided by submitting the full name of a person. However, for our evaluation we prefer to have, if possible, a mention of the occupation in the text. Thus, the format of the queries used to retrieve documents for our corpus is:

\[\text{\langle person\rangle}\text{Last_Name}/\text{\langle person\rangle}\text{\langle role\rangle}\text{Occupation}/\text{\langle role\rangle}\text{.}\]

The number of documents retrieved varied from one, for the query

\[\text{\langle person\rangle}\text{Cauchy}/\text{\langle person\rangle}\text{\langle role\rangle}\text{mathematician}/\text{\langle role\rangle}\text{.}\]

to up to 8144, for

\[\text{\langle person\rangle}\text{Clinton}/\text{\langle person\rangle}\text{\langle role\rangle}\text{politician}/\text{\langle role\rangle}\text{.}\]

To counteract misbalance in the data we relied on the tf-idf ranking of JuruXML to sort the matching documents. The top ten such documents were kept (or all of them if fewer than ten were returned).

4. Extracting biographical activities

We assume that most biographies can be broken into three main parts: biographical information, information describing activities of the person with respect to his/her occupation, and person-specific activities. Biographies can be created according to the occupation. Though there are some general facts typical for any biography, such as: date and place of birth,
places where person lived, etc. (all these facts are gathered in one bio feature in Zhou et al. [55]), people of different occupations get famous for performing activities within their occupations. Singers sing, explorers travel around the globe discovering and studying new lands, artists create paintings, etc.

To automatically discover general and occupation-related activities we use a modified version of atomic relations/events described in Filatova and Hatzivassiloglou [21]. Similar technique for relation discovery is used in Hasegawa et al. [29]. Atomic relations are triplets consisting of two named entities and a verb or an action-defining noun which labels the relation between these two named entities. Verbs are considered to be central elements for analyzing relations among words in text for many NLP applications [31,35,36,41,45].

We extract 189 lists of atomic relations (a list of atomic relations for each person) according to the following procedure:

1. For each person analyze the corresponding collection of documents retrieved for this person.
2. From every sentence containing the name of the person under analysis extract all the pairs of named entities, one of the elements of which is the name of this person.
3. For every such pair of named entities extract all verbs, excluding modal and auxiliary verbs, that appear in-between them, as we are interested only in the activities performed by people.
4. Count how many times each atomic relation containing two named entities and a verb in-between appears in the collection of documents describing the person under analysis.

The NE tagger we use is a derivative of that described in Prager et al. [43]. It tags named entities of about 100 types arranged in an ontology. Some of the marked types are very specific, like ZIPCODE and ROYALTY. To avoid overfitting we choose six high-level types for atomic relations extraction: PERSON, PLACE, DATE, WHOLENO, ORG and ROLE.

In this work, we do not use syntactic dependencies among words in a sentence. Rather we rely on shallow co-occurrence statistics. We use this co-occurrence information to model a biography structure for several occupations. Filatova et al. [22] analyze syntactic dependencies to model the structure of several events: earthquakes, terrorist attacks, plain crashes, and presidential elections. Table 3 contains examples from the list of atomic relations extracted for Christopher Columbus.

We analyze all the atomic relations extracted for different people belonging to the same occupation. A description of each person is a separate collection of documents. Thus, we cannot use the original atomic relation scores described in Filatova and Hatzivassiloglou [21]; instead, we keep simple counts for the triplets. Later, we combine triplets extracted for different people. This technique is similar to the one used for domain template induction where each domain instance description was represented by a separate document collection [22].

The task of generalized event extraction can be cast as the task of Open or On-Demand Information Extraction (OIE), or Unrestricted Relation Discovery [7,29,48,49]. OIE deals with creating data repositories for all possible relations present in the input corpus. In contrast to classic IE systems [13,17,26], OIE systems do not require as input a pre-defined set of relations. Rather, OIE systems attempt to capture all important relations mentioned in the input document or collection of documents. Generalization of atomic relations is also similar to data anonymization [37].

4.1. Generalized atomic relations

Our goal is to collect information about activities general for all people and about activities specific to some occupations. Thus, we are interested in the semantic information about the extracted atomic relations rather than in the relation between the exact named entities encoded by each relation. Thus, we analyze not the atomic relations themselves but the generalized versions of the extracted atomic relations.

The generalized atomic relations identify to what types of named entities people of various occupations are linked through the activities (verb and action relation nouns) they perform. For example, here are two sentences about explorers:

Amerigo Vespucci explored the shores of South America.
Vitus Bering explored the Aleutian Islands.

The corresponding atomic relations extracted for these sentences are presented in Table 4. Clearly, the atomic relations presented in Table 4 capture information about the same type of activity, namely that explorers explore various locations. Thus, these two atomic relations describe the same activity. What make these atomic relations different is the exact names of the explorers and the locations a particular explorer explored. We can unify these atomic relations by omitting the exact named entities and leaving only their types. The resulting atomic relations we call generalized atomic relations. In the generalized atomic relations we distinguish two types of named entities with the tag PERSON: those which refer to the person under analysis (from

<table>
<thead>
<tr>
<th>First named entity</th>
<th>Verb</th>
<th>Second named entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columbus/PERSN</td>
<td>Died/VBN</td>
<td>1506/DATE</td>
</tr>
<tr>
<td>Columbus/PERSN</td>
<td>Sailed/VBD</td>
<td>India/PLACE</td>
</tr>
</tbody>
</table>

Table 3
A sample of atomic relations extracted for the collection of documents about Christopher Columbus.
now on they are marked as NAME/PERSON) and all the rest (marked as PERSON). Thus, we separate the person under analysis from the people who are linked to this person through various activities. The atomic relations presented in Table 4 can be converged to the following generalized atomic relation:

NAME/PERSON \rightarrow \text{explored}/VBD \rightarrow \text{PLACE}.

In this work we use generalized atomic events. This generalization technique is similar to the one used by researchers for semantic pattern discovery for information extraction [8,15,50,53]. Filatova and Hatzivassiloglou [21] show that atomic relations can be used to capture the most important relations described in a text collection and assign to them good-quality labels. In this work we show that generalized atomic relations can be used for identifying occupation-related activities. These activities can be used for describing people’s life taking into consideration their occupations.

5. Identifying occupation-related activities

We assume that the activities important for an occupation are linked to the named entities important for this occupation and vice versa; the named entities important for this occupation are linked to the representatives of this occupation through the important activities. Formulated like this, the problem of identifying the actions important for an occupation can be solved using the methodology suggested by Kleinberg [33] for ranking web-sites, where a search engine counts “inbound and outbound links to identify central sites in a community.”

The major idea of this technique is based on the assumption that good hubs contain links to good authorities and that links to good authorities are listed within good hubs. Treating activity verbs as hubs and named entity tags as authorities, we map the problem of scoring the reliability of web pages as it is formulated for search engines using the Hubs and Authorities (HITS) technique Kleinberg [33] for scoring the reliability of web pages.

### Table 4

A sample of atomic relations extracted for two representatives of the explorer occupation.

<table>
<thead>
<tr>
<th>First named entity</th>
<th>Verb</th>
<th>Second named entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amerigo Vespucci/PERSON</td>
<td>Explored/VBD</td>
<td>South America/PLACE</td>
</tr>
<tr>
<td>Bering/PERSON</td>
<td>Explored/VBD</td>
<td>Aleutian/PLACE</td>
</tr>
</tbody>
</table>

To select occupation-related activities we merge lists of generalized atomic relations corresponding to the people of the same occupation. Hence, we get ten lists of generalized atomic relations corresponding to the ten occupations under analysis.

The count for a generalized relation for a specific person is the sum of all the counts of the atomic relations that converged to this generalized relation. After that, we create a list of generalized atomic relations for every occupation. A count for an occupation-generalized event is the sum of all the counts of person-generalized relations.

We assume that the named entities which are important to the people of some occupation are connected to those people through the activities which are important for the people of this occupation. After formulating the problem of identifying occupation-related activities in terms of connecting important named entities through important activities, we can cast this problem and the problem of scoring the reliability of web pages as it is formulated for search engines using the Hubs and Authorities (HITS) technique Kleinberg [33] for scoring the reliability of web pages.

Variables $m$ and $k$ are unique for every occupation.

To avoid data sparseness we use a smoothing factor $c = 0.01$. The elements of the $P_{N \rightarrow V}$ matrix are defined in Eq. (1).

$$P_{N \rightarrow V}[i,j] = \frac{f(n_i \rightarrow v_j)}{\sum_{v \in V} f(n_i \rightarrow v)} + c$$  (1)
where \( f \) is equal to the sum of the counts of all the generalized atomic relations containing this link for the occupation under analysis. In the same way we define \( P_{V \rightarrow N} \) \( k \times m \) row-stochastic transition matrix from verbs to named entities (Eq. (2)):

\[
P_{V \rightarrow N|i,j|} = p_{v_i,n_j} = (1 - c) \frac{f(v_i \rightarrow n_j)}{\sum_{n \in \mathcal{N}} f(v_i \rightarrow n)} + c.
\]

Using \( P_{V \rightarrow N} \) and \( P_{N \rightarrow V} \), we can define the square stochastic \( P_{V \rightarrow V} \) transition matrix (Eq. (3)):

\[
P_{V \rightarrow V} = P_{V \rightarrow N} \cdot P_{N \rightarrow V}
\]

that can be used for scoring the verbs according to how important they are for the current occupation. Due to the construction rules, this matrix is a square stochastic matrix where the sum of all the elements in each row and each column is equal to one and can be analyzed in terms of Markov Chains.

According to Markov Chain Theory [32], for a square stochastic matrix it is possible to find a steady state which corresponds to the eigenvector for the eigenvalue equal to 1. Any square stochastic matrix has 1 among its eigenvalues. The length of the steady state vector is then adjusted so that the sum of all elements of the vector is equal to 1.

The same way the eigenvector corresponding to the steady state for web-pages ranks these pages, the eigenvector corresponding to the steady state of transition matrix described in Eq. (3) ranks how tightly the activities are linked to the occupation under consideration.

We rank the extracted activities in the same way web-pages are ranked in Hubs and Authorities, using the steady state vector. The higher the value of the activity, the higher the chances for this activity to be used for the description of a person corresponding to the occupation for which the transition matrix described in Eq. (3) is created.

The dimensionality of this matrix depends on the variety of the verbs in all forms used in the generalized atomic relations for the representatives of this occupation and varies from 800 for physicists up to 2100 for politicians in our data. This transition matrix models the flow of an imaginary text. For this imaginary text, a description of the activity which should be output next depends on the previously output activity. After large enough amount of iteration the knowledge of the last output activity becomes unnecessary and the probabilities of outputting information about every activity stabilize.

After each activity gets such a score we can choose those activities which, separated from the rest of the activities, can potentially name all the major activities that are usually typical for the people of the current occupation. The eigenvector corresponding to the steady state of the matrix contains information about how tightly each activity included in the matrix is related to the current occupation.

Table 5 contains top ten activities for four occupations: artists, dancers, physicists and singers. These activities are listed in the sorted order, the ones on top of the table have the highest scores in the respective steady state vectors.

The activities presented in Table 5 can be divided into three types:

- those which are occupation-related, such as danced and perform for dancers, discovered for physicists, singing and sang for singers;
- those which are likely to be used in any biography, such as born, died, became;
- other, which are mostly general purpose, modal, or auxiliary verbs, such as been, made.

The goal of our classification task is to assign a person to her respective occupation. Thus, for our classification we rely mainly on the first type of activities, i.e. those activities that capture the description of a person as a practitioner of a particular occupation. To extract the activities of the second type we create \( P_{V \rightarrow V} \) a transition matrix for the combined set for all the generalized atomic relations created for all the ten occupations. According to the matrix construction rules that we describe
above, this matrix is also stochastic and by calculating the eigenvector corresponding to its steady state we can identify those activities which are tightly linked to any person irrespectively of his/her occupation and thus reflect general biographical information. Table 6 contains top ten activities for this matrix. Clearly, Table 6 contains verbs corresponding to three types of activities: general biographical, occupation-related and person specific.

In Section 6 we show that the lists of occupation-related and general activities are reliable features for classifying people according to their occupations.

6. Classification

In this section we describe people classification according to their occupations by exploiting the information obtained from text according to the procedure described above.

Choosing the appropriate classifier requires considering several aspects of the classification task [16]. For our classification experiments we use a multi-class SVM classifier that is one of the most widely used classifier for a variety of NLP tasks. As we have 189 data-points corresponding to ten classes we use leave-one-out cross validation which allows us to use the maximal possible amount of data for training. We experiment with two sets of features: one set consists only of the verbs corresponding to the occupation-related activities (Section 6.1); the other set consists of the complete generalized atomic relations (Section 6.2).

In this section we demonstrate the advantage of using generalized atomic relations in comparison to using verbs from these relations.

6.1. SVM classification using only verbs

To get verb features for multi-class SVM classification we use ten occupation-related lists of activities. Each of these lists is a sorted list of verbs for a respective occupation. The verbs are sorted according to the respective values of the eigenvector corresponding to the steady state of the respective occupation matrix.

The verb-only algorithm is as follows:

V1 Get the sorted list of activities (verbs) for every occupation (ten lists). These activities are the major features on which SVM relies to assign an occupation to a person.
V2 Get the sorted list of activities for all occupations merged together. These activities are used in Step V4 to remove from the list of classification features those activities which are general and not helpful for identifying the occupation of a person.
V3 Get the top 15% of the activities from each of the ten occupation-related lists and the list of general activities.
V4 From the ten occupation-related lists remove those activities which are also present in the list of general activities.
V5 Merge ten occupation-related lists into one list and remove from this list all the activities that appear in more than 2 occupations.

By leaving at Step 3 only some percentage of the activities (verbs) instead of an absolute amount, we take into account the fact that the number of activities used to describe different occupations varies greatly from occupation to occupation (for example, 1794 activities are used in the atomic relations for composers, and only 800 for physicists).

As the activities get scores according to the steady state vectors, the activities with high scores are the ones which are most likely to be used for the description of a person of the current occupation. The activities with low values are too specific and are likely to be used in only a few descriptions of people of this occupation. For example, for composers we want to keep all the activities related to composing music. We know that Alexander Borodin was both a composer and a chemist: we do not want to keep those specific verbs which describe his activities as a chemist in the list of the activities describing composers.

In Step 4 we remove from our final list those activities that are typical for all humans and thus cannot be used to distinguish among different occupations. In Step 5 we make our activities as specific as possible: For example, there will be some intersection in activities among mathematicians and physicists, and such activities cannot be helpful for differentiation between these occupations.

The final activities list is used as the list of features for SVM classification. Then we assign values to these features for every person: if the activity from the features list is used as a connector for the extracted atomic relations, then this feature receives the value of 1, if there is no atomic relation using this activity as a connector then this feature is assigned 0. We use binary values for our features instead of the atomic relation counts because the reliability of the scores for the atomic relations extracted for different people varies greatly. For some people we retrieve 10 documents with many biographical facts about those people, for other people we retrieve 2 or 3 documents which only mention the people queried.

Selection of the appropriate set of features for the classifier is an important step toward getting high quality results [39]. Removal of some of the features is reinforced by the Koller and Sahami [34] work on feature selection for document classification. It indicates that keeping only a small fraction of the available features improves the classification performance. The optimal
number of features is still to be determined. For our classification experiments we concentrate on using occupation-related activities only.

We train our classifier and evaluate its performance using leave-one-out cross validation. Out of 189 people, eight are not assigned any features. This is because all the atomic relations extracted for these 8 people are either too general or too specific. As these 8 people do not have any features to assign them to the most likely occupation, they are misclassified to the default occupation (artists). Six of these eight are mathematicians, one is a dancer, one is a physicist. Absence of verbal features can be explained by the small number of the documents retrieved for these people. In fact, three or fewer documents were retrieved for these people. Table 7 shows how many documents are analyzed per person on average for each occupation (column “Average # of Docs”), as well as how many practitioner for every occupation are there in our corpus (column “# of Reps”). Due to the nature of our document collections, the smallest number of documents analyzed was for mathematicians and physicists: current newswire texts (AQUAINT and TREC corpora) do not contain much information about scientists, and those parts of the World Book and Encyclopedia Britannica which we had at our disposal only contained information for some of the scientists. Out of the remaining 181 people, only 90 are classified correctly.

We believe that the performance of SVM classification based solely on verbs is poor because it does not take into account the information that many activities which are expressed with the help of the same verb are surrounded by different types of arguments for different occupations. For example, Henri Matisse is classified as a dancer based on the frequent co-occurrence with the dance activity, which is understandable as one of his most famous paintings is “Dance”. Or, the explored activity is among the top activities for several occupations: writers, composers, explorers; but this activity is linked to the PLACE named entity tag only for the explorers.

Though the classification based solely on verbs gives quite poor results we consider it to be a valid starting point for our classification experiment as usually activities are associated with the verbs corresponding to these activities. Next, we show that the SVM classification with generalized atomic relations as features significantly outperforms the SVM classification using solely verbs.

6.2. SVM classification using atomic relations

To create generalized atomic relation features for multi-class SVM classification we use the sorted lists of activities for the ten occupations and the general list of activities. The activities are sorted according to the values they get from the eigenvector corresponding to the steady state.

AR1 Same as step V1 above.
AR2 Same as step V2 above.

Table 5
Top ten activities for three occupations: dancers, physicists and singers.

<table>
<thead>
<tr>
<th>Artist</th>
<th>Dancer</th>
<th>Physicist</th>
<th>Singer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Born/VBN</td>
<td>Made/VBD</td>
<td>Born/VBN</td>
<td>Said/VBD</td>
</tr>
<tr>
<td>Painted/VBD</td>
<td>Died/VBD</td>
<td>Died/VBD</td>
<td>Born/VBN</td>
</tr>
<tr>
<td>Painted/VBN</td>
<td>Appeared/VBD</td>
<td>Announced/VBD</td>
<td>Died/VBD</td>
</tr>
<tr>
<td>Including/VBG</td>
<td>Be/VB</td>
<td>Discovered/VBD</td>
<td>Died/VBD</td>
</tr>
<tr>
<td>Became/VBD</td>
<td>Became/VBD</td>
<td>Be/VB</td>
<td>Singing/VBG</td>
</tr>
<tr>
<td>Died/VBD</td>
<td>Born/VBN</td>
<td>Became/VBD</td>
<td>Has/VBZ</td>
</tr>
<tr>
<td>Been/VBN</td>
<td>Danced/VBD</td>
<td>Wrote/VBD</td>
<td>Conducting/VBG</td>
</tr>
<tr>
<td>Showed/VBD</td>
<td>Blessed/VBN</td>
<td>Helped/VBD</td>
<td>Made/VBD</td>
</tr>
<tr>
<td>Had/VBD</td>
<td>Perform/VB</td>
<td>Named/VBN</td>
<td>Became/VBD</td>
</tr>
</tbody>
</table>

Table 6
Top ten activities common for all the eleven occupations.

<table>
<thead>
<tr>
<th>Biography-related verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Said/VBD</td>
</tr>
<tr>
<td>Born/VBN</td>
</tr>
<tr>
<td>Died/VBD</td>
</tr>
<tr>
<td>Wrote/VBD</td>
</tr>
<tr>
<td>Became/VBD</td>
</tr>
<tr>
<td>Had/VBD</td>
</tr>
<tr>
<td>Known/VBN</td>
</tr>
<tr>
<td>Be/VB</td>
</tr>
<tr>
<td>Included/VBD</td>
</tr>
<tr>
<td>Including/VBG</td>
</tr>
</tbody>
</table>
AR3 For the top 15% of the activities from the ten occupation-related lists get all the generalized atomic relations containing those activities.

AR4 For the top 15% of the activities typical for all the occupation (Step AR2) get all the generalized atomic relations containing those activities.

AR5 From the ten occupation-related lists (Step AR3) remove those generalized atomic relations which are also present in the list of general generalized atomic relations (Step AR4).

AR6 Merge the ten occupation-related lists into one and remove from it all the generalized atomic relations that appear in more than 2 occupations.

Out of 189 people, nine are not assigned any features. This again is because all the atomic relations extracted for these 9 people were either too general or too specific. The people who do not get any relation features are the same as those who do not get any verb features plus one physicist.

As these nine people did not have any features according to which they could be assigned to the most likely occupation, they were assigned to the default occupation, which means that they are misclassified (6 of the people that did not get any features for classification were mathematicians, 1 person was a physicist and 1 person a dancer). Even despite the fact that the initial starting point is much worse than for the previous experiment the performance of SVM classification using generalized atomic events as features is much better than SVM classification using solely verbs as features.

Out of the remaining 180 people, 124 people are classified into the appropriate occupations. Table 7 shows that generalized atomic relations are more reliable for occupation classification than plain activities extracted from these generalized atomic relations. Thus, it can be concluded that structured information captured by atomic relations is valuable and reliable. For example, using atomic relations, Henri Matisse is now correctly classified as an artist.

According to the t-test the performance of the classification based on atomic relations is significantly better \((p<0.05)\) than the performance of the classification based solely on activities.

We would like to note that after closer analysis some of the cases of misclassification can be considered as correct assignments as a person could excel in different occupations. For example, in our corpus Paul McCartney is defined as a singer, but classifying him as a composer is a valid result as well.

### 6.3. Other types of classification

The task of classifying people according to their occupations is new and to our knowledge there is no standard baseline against which we could compare our results with. Thus, we adapt for comparison two classification techniques used for other tasks: random assignment of an occupation and mutual information for a name of a person co-occurring with a title an occupation.

#### 6.3.1. Random occupation assignment

As the distribution of people among the occupations in our corpus is unknown we cannot give a single exact probability of assigning a correct occupation to a particular person. Instead, a fair rough approximation to this probability is calculated.
according to Eq. (4). The results are presented in Table 7. According to these results, SVM classification is better than a random assignment of an occupation even for the case where only verbs are used as features for SVM classification.

\[
P_{\text{correct occupation}} = \frac{1}{10} = 0.1.
\]  

Random assignment gives a very low baseline which we easily outperform. For this reason, we use another classification based on mutual information to estimate how good our results are.

6.3.2. Mutual information (MI)

For this classification we use the same four document collections from which we retrieve documents about the people in our corpus (namely, AQUAINT, TREC, part of World Book and part of Encyclopedia Britannica). However, we use the complete four document collections rather than only the documents retrieved for atomic relation identification.

First, we get the counts of how many documents are retrieved for the queries containing only the titles of the occupations (e.g., “(role) mathematician/(role)”, “(role) artist/(role)” etc.). Then, we get the counts for the queries containing all possible combinations of people's names and occupations' titles. For example, “(role) mathematician/(role)(person) Picasso/(person)”, “(role) artist/(role)(person) Picasso/(person)”, etc. Finally, we divide the counts for the queries submitted for the occupation plus person by the count for the corresponding occupation query. The maximum of all the ratios for the person gives the occupation for this person (Eq. (5)).

\[
\text{Occupation}_j = \max_{\text{for all } i} \frac{\text{count}_{\text{occupation}_i, \text{person}_j}}{\text{count}_{\text{occupation}_i}}.
\]  

According to Table 7 SVM, classification based on atomic relations outperforms MI classification for six occupations out of ten, for one occupation (explorers) the results for SVM classification and MI are the same and for three occupations MI classification outperforms SVM classification. One of the cases where MI classification outperforms SVM classification is mathematicians, where nine mathematicians have no features in SVM classification and thus, do not have any better than random chance to be classified correctly.

Thus, our SVM classification of people according to their occupations based on atomic relations has performance comparable to MI-based classification. This is significant since for many NLP-related applications, MI classification outperforms other methods and is considered to be one of the most powerful classification methods [30,40].

6.4. Occupation classification using web as a corpus

The text corpus used for the classification discussed above consists of four corpora: AQUAINT, TREC, part of World Book and part of Encyclopedia Britannica. The documents in these corpora are preprocessed and include XML tags identifying different parts of the documents (i.e., title, date, text body, etc.). The JuruXML search engine supports the use of tagged terms in the search queries.

To study how well the presented methodology for biographical activities identification can be generalized we run the classification experiment using web as a corpus.

6.4.1. Data collection

Given an infinite number of web page formats, automatic extraction of text body from an arbitrary page is a very difficult task [25,28] that is outside the scope of this paper. To avoid the problem of dealing with unprocessed HTML text we employ human subjects to provide us with the text blocks from the related web pages. To do this we use the Mechanical Turk service\(^8\) that allows to “crowdsource” labor intensive tasks and is now being used as a source of subjects for experimental research [3,19,42].

For example, to get text documents about Alicia Alonso we asked MTurkers to: find 10 web pages about Alicia Alonso and submit the URL and the plain text extracted from the page:

- The submitted web pages should report facts and stories about Alicia Alonso's life and professional activities (occupation of Alicia Alonso is dancer);
- The plain text that you extract from a web page should contain the actual text content of the page (and not ads, navigation links, etc.);
- The plain text should contain at least three sentences;   
- All 10 URLs should be distinct.

We asked to submit information from 10 web pages as we wanted to avoid misbalance in the number of documents for different people. Also, the goal of this experiment is to compare the results obtained from the web documents against the results

---

\(^8\) https://www.mturk.com.
from the corpus consisting of four tagged corpora: AQUAINT, TREC, part of World Book and part of Encyclopedia Britannica retrieved by JuruXML.

To avoid the problem of employing spammers who could potentially submit random URL links and/or text content extracted from random web pages, we first ran a preliminary data collection experiment for a small number of people in our corpus to find annotators who submitted valid URL links as well as valid text extracted from the corresponding web pages. While checking for the correctness of the submitted web pages we were intersected in any pages mentioning a person. The content of those pages varied greatly: from Wikipedia biographies of the people in our corpus to news articles that mentioned these people only a couple of times. Thus, any page that mentioned a person from our corpus was considered valid. After running the preliminary experiment, we created a list of MTurkers who submitted valid links and extracted valid text content. We then ran the data collection experiment using only those MTurkers.

After creating a corpus of 1890 documents (10 documents for each of the 189 people in our corpus) we created the atomic relation and verb classification features and ran an SVM classifier following the procedure described in Sections 5 and 6. The classification results for these experiments are presented in the bottom part of Table 7.

6.4.2. Baseline classification for the web corpus

The random baseline does not depend on the corpus.

6.4.2.1. Mutual information baseline. To run the mutual information baseline we used the Microsoft Bing Search Engine. As Bing.com does not allow the presence of named entity tags in the queries we could not run the MI baseline experiment identical to the one reported in Table 7. Rather, we decided to use two types of queries: one—with only the last name of a person and a candidate occupation; the other one—with the full name of a person and a candidate occupation; plus the queries containing only the name of the occupation:

- + last name + occupation + Alonso + artist
  + Alonso + athlete
  + Alonso + composer
  + Alonso + dancer
  ...
- + full name + occupation + "Alicia Alonso" + artist
  + "Alicia Alonso" + athlete
  + "Alicia Alonso" + composer
  + "Alicia Alonso" + dancer
  ...
- occupation + artist
  + athlete
  + composer
  + dancer
  ...

To compute the Mutual Information prediction of the occupation of a person we use Eq. (5). The classification results for these two MI baseline experiments are presented in Table 7.

The MI results for the web corpus presented in Table 7 are predictable: the more precisely we define the person—the higher the chances that we correctly predict the occupation of this person. The SVM classification results for both verbs and atomic relations are better than the results with the MI baseline using the last name only. The MI baseline that uses the full name of a person is a very strong baseline and we do not outperform it.

6.4.2.2. Unigrams baseline. Using the constructed web corpus we introduce another bag-of-words or unigram baseline that uses all the words from the document collections about the people belonging to a particular occupation as binary features to predict an occupation of a new person. Such a unigram baseline is frequently used in NLP classification tasks. The classification results for the unigram baseline experiment are presented in Table 7. The SVM classification results for both verbs and atomic relations are better than the SVM classification results based on unigrams.10

---

9 http://www.bing.com/

10 When using unigrams as features, we also experimented with different feature selection algorithms [1,2]. We used the HITON_PC and HITON_MB algorithms, trying various settings. The classification performance remained relatively unchanged when using unigrams as features and a variety of possible classification algorithms. The maximum observed classification performance was 0.580, measured using cross-validation. This is only a small increase compared to the baseline case that had an accuracy of 0.55; we should also mention that often the performance was lower than the baseline case of no feature selection at all.
7. Classifications' results discussion

Though we do not outperform the MI baseline for the initial corpus and the MI baseline using the full name of a person for the web corpus, our methodology has one crucial advantage. We use classification not as a primary task but as an evaluation testbed; we show that the lists of generalized atomic relations created for every occupation and for general biographies indeed capture the major activities performed by people of the respective occupations and can be used for biography generation. For example, the generalized atomic relations which are used for the description of representatives within all the ten occupations and are excluded from the list of features for SVM classification as too general, contain verbs such as born/VBN, died/VBD linked to the DATE and PLACE named entity tags or became/VBD linked to the ROLE named entity tag. Table 8, by contrast, contains occupation-related generalized atomic relations. These generalized atomic relations have high scores within the respective occupations, are used as features for SVM classification and have non-zero values in the feature sets which correctly classified people into the appropriate occupations.

Table 9 contains a sample of generalized atomic relations which were used for the representatives of all the ten occupations analyzed and thus can be used for either identifying the snippets of text which contain biographical information or they can be used for constructing auxiliary questions.

In this section, we have shown that the occupation-related activities learned from descriptions of sets of people belonging to a pre-defined set of occupations can be successfully used as classification features to assign a new person to her respective occupation from the initial occupation list. Moreover, using the same methodology we can learn general biographical activities that are likely to be used in a description of any person irrespective of occupation.

8. Clustering

By definition, a classification task requires a pre-defined set of classes into which the input elements should be classified. In most cases, including human occupations, it is not possible to come up with an exhaustive list of such classes. Thus, many tasks use clustering rather than classification to group closely related elements. In this section, we describe how clustering can be used for grouping people according to the activities used in the descriptions of these people. We also show that clustering can be used not only for identifying activities typical for some particular occupations but also for identifying which activities can be used to describe people of several occupations and, propagating further, which activities are general biographical ones. Thus, in contrast to the classification experiment, we do not eliminate general biographical activities from the features list. Rather, we use all generalized activities identified in the people’s description and show that general biographical and occupation-related activities can be identified automatically as a side-effect of the clustering process.

The generated hierarchy of clusters links those occupations that have intersecting sets of occupation-related activities. This hierarchy contains information about various human occupations together with practitioners of each occupation and a set of activities which are used in the descriptions of people belonging to each particular occupation. The root node of this hierarchy contains activities which correspond to general biographical information and can be used in a biography of any person. Thus, the final hierarchy, which is learned in an unsupervised fashion from unstructured data (text), can be used as an initial step for building an ontology of human occupations. At the same time, the described methodology can be potentially used to tease apart descriptions of people having the same name but belonging to different occupations, as in Artiles et al. [6].

For our clustering experiments, we use the generalized atomic relations obtained from the corpus described in Section 3.

8.1. CLUTO clustering tool

For our experiments, we used the CLUTO clustering toolkit which has been shown to produce high-quality clustering solutions [54]. In our experiments, we used k-way clustering with a direct clustering procedure when all k clusters are formed simultaneously. We use CLUTO I2 criterion function which is the default clustering criterion function for k-way clustering. I2 criterion function is computed according to Eq. (6):

\[ I_2 = \sqrt{ \frac{1}{k} \sum_{i=1}^{k} \sum_{u \in S_i} \sum_{v \in S_j} \text{sim}(u, v) } \]

where \( k \) is the total number of clusters, \( S \) is the total number of objects to be clustered (in our case, \( S = 189 \)—the number of people in our corpus), \( u \) and \( v \) represent two objects from \( S \), \( S_i \) is the set of objects assigned to the \( i \)th cluster, \( \text{sim}(u, v) \) is the similarity function between two objects (in our case cosine similarity between two vectors corresponding to two people).

---

11 The generalized activities used in the clustering experiments are extracted from the corpus that consists of four corpora: AQUAINT, TREC, part of World Book and part of Encyclopedia Britannica.
We use the activities extracted in Section 4 as features for the clustering experiments. This data was encoded in a $N \times M$ matrix, where $N$ was the number of people to be clustered (189) and $M$ was the number of unique activities (generalized atomic events) used in the description of all the 189 people ($M$ is equal to 21,433). The elements of the $N \times M$ matrix are assigned 0 or 1 according to the following criterion:

$$a_{ij} = \begin{cases} 1, & \text{if the activity } j \text{ is used in the description of the person } i; \\ 0, & \text{otherwise.} \end{cases}$$

We used binary values for our features instead of statistics about the activities for each person because the document collections retrieved differed in sizes and quality.

### 8.2. Identifying the number of clusters

Any clustering algorithm assumes the knowledge of a pre-defined parameter which either defines how close the elements of the clusters are or the number of clusters that should be created. We ran several experiments grouping people into different number of clusters. We show, that when the number of clusters is chosen to be equal to the number of the underlying occupations, then clustering can be used for automatic grouping of people belonging to the same occupation. We also show, when the number of outcome clusters is smaller than the number of the underlying occupations, how the result of clustering can be used for grouping occupations having similar activities. We also compare the grouping of people inside clusters against the gold standard occupation assignment.

<table>
<thead>
<tr>
<th>Table 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupation-related generalized atomic relations obtained from the corpus used for the first set of experiments (note that NAME stands for NAME/PERSON). The presented atomic relations are extracted from the corpus that consists of four corpora: AQUAINT, TREC, part of World Book and part of Encyclopedia Britannica.</td>
</tr>
<tr>
<td>Artists</td>
</tr>
<tr>
<td>NAME - painted/VBN - DATE</td>
</tr>
<tr>
<td>NAME - resemble/VB - PERSON</td>
</tr>
<tr>
<td>PERSON - designed/VBN - NAME</td>
</tr>
<tr>
<td>Composers</td>
</tr>
<tr>
<td>NAME - composed/VBN - PERSON</td>
</tr>
<tr>
<td>NAME - include/VBP - WHOLENO</td>
</tr>
<tr>
<td>ROLE - hearing/VBG - NAME</td>
</tr>
<tr>
<td>Explorers</td>
</tr>
<tr>
<td>NAME - annexes/VBZ - PLACE</td>
</tr>
<tr>
<td>NAME - reach/VB - PLACE</td>
</tr>
<tr>
<td>NAME - declares/VBZ - DATE</td>
</tr>
<tr>
<td>Physicists</td>
</tr>
<tr>
<td>DATE - described/VBD - NAME</td>
</tr>
<tr>
<td>ROLE - predicted/VBD - NAME</td>
</tr>
<tr>
<td>NAME - continued/VBD - ORG</td>
</tr>
<tr>
<td>Singers</td>
</tr>
<tr>
<td>NAME - conducting/VBG - PLACE</td>
</tr>
<tr>
<td>NAME - sing/VB - ROLE</td>
</tr>
<tr>
<td>NAME - sang/VBD - PERSON</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalized atomic events typical for any biography. The presented atomic relations are extracted from the corpus that consists of four corpora: AQUAINT, TREC, part of World Book and part of Encyclopedia Britannica.</td>
</tr>
<tr>
<td>First named entity</td>
</tr>
<tr>
<td>NAME/PERSON</td>
</tr>
<tr>
<td>NAME/PERSON</td>
</tr>
<tr>
<td>NAME/PERSON</td>
</tr>
</tbody>
</table>
8.2.1. Ten clusters

Identifying the optimal number of clusters into which the input data should be broken is a difficult task. We do not try to solve this problem as it is outside of the scope of this paper. In our corpus there are representatives of ten occupations. Some of these occupations are very distinct from the rest of the occupations and are not likely to have a lot of occupation-related activities used by “outsiders” (e.g. compare dancers and explorers). On the other hand, some occupations are quite close and can have substantial intersections of the lists of occupation-related activities (e.g., physicists and mathematicians). This issue complicates the task of choosing the optimal number of clusters.

For our initial experiment, we assume that the optimal number of clusters corresponds to the number of occupations covered in our corpus. We also assume that we know in advance how many occupations are covered in our corpus. Thus, for the initial clustering experiment, the input number of clusters into which people should be divided is equal to ten.

Fig. 3 contains the result of running the clustering algorithm with ten as a parameter identifying the number of clusters into which the input elements (people) should be divided. In Fig. 3 we ordered the clusters in such a way so that the diagonal is maximized (i.e., the row where ith element is maximal is placed in the ith row). The identification numbers for clusters presented in the cid column are the same as they are produced by the CLUTO clustering tool, where “clusters that are tight and far away from the rest of the objects have smaller cid values.” This manual re-ordering of clusters is performed solely for the visualization purpose and it does not modify any of the numbers computed by the clustering procedure. The tightness of clusters is calculated as difference between internal and external similarities, where internal similarity is the average similarity between the objects in the cluster and external similarity is the average similarity of the objects of some cluster and the rest of the objects from the other clusters.

Fig. 4. Grouping people into two clusters.
One can see that all the created clusters except one (Cluster 2) have a dominant occupation which can be used as a label for the respective cluster. For example, 14 out of 18 members of Cluster 6 are artists. For Cluster 2, whose elements are spread among several occupations, it is not possible to pinpoint one dominant occupation.

The same can be noticed about occupations. For most occupations it is easy to point out the cluster containing the vast majority of its representatives. For example, 16 out of 17 explorers are placed into Cluster 5. Only writers are spread among five clusters, with no cluster distinctively standing out as having the vast majority of writers.

Using CLUTO it is possible to analyze the distribution of clustering features, and to investigate which activities are the most descriptive and most discriminative ones. In our case, we are interested in analyzing which activities are related to which clusters.

8.2.2. Two clusters

It must be noted that the problem of estimating the optimal number of clusters for a data set is one of the most essential issues in cluster analysis. We do not tackle this problem in this work. Rather, we show that for the problem of clustering people according to their occupations even a non-optimal number of clusters gives meaningful and useful results.

Fig. 4 shows the distribution of people between two clusters, when k is set to 2. In Fig. 4 the separation between explorers, mathematicians and physicists on the one side, and representatives of other occupations on the other side is clear-cut. This distribution can be roughly mapped to the division of the occupations used in our experiments into Arts and Sciences.

9. Conclusions and future work

In this paper, we explore the hypothesis that the descriptions of people belonging to the same occupations are likely to contain similar sets of activities. This hypothesis leads toward the discovery of the hierarchical structure of biographical activities. For example, some activities are typical for most people irrespective of their occupation, while other activities are typical for people of a particular occupation.

The contributions of our paper are three-fold:

- We introduce a novel representation (generalized atomic relations) for describing human activities and show how random walk theory can be used for ranking the extracted activities according to their importance within three categories: general biographical, occupation-related, and person-specific activities.
- We use the extracted activities features for SVM classification and show that the results of this classification outperform several widely-used classification baselines. Moreover, the classification features have an add-on occupation descriptive value.
- We use generalized atomic relations as clustering features. We show that irrespective of the input number of clusters, the results are meaningful. This is an interesting clustering side effect that can be used for automatic generation of an ontology describing the relation among various occupations.

In the future, we are interested in investigating ways of identifying other types of activities which are neither general nor occupation-related but rather person-specific. We believe that the use of generalized atomic relations can enable significant new techniques for a number of natural language processing tasks (i.e., QA, biographical summarization, IR query expansion, etc.). We also plan to experiment with learning within the classification stage to contrast different occupations and people within the respective occupations [52].

Also, we are interested in automatic ontology creation. We believe the clustering experiment described in Section 8 can be used as a starting point for creating a hierarchy of occupations. We believe that analyzing descriptions of people belonging to various occupations, we can build a hierarchy of occupations. Nodes of this hierarchy correspond to specific occupations of groups of related occupations. Also, each node of this hierarchy lists practitioners of these occupations as well as activities typical to these occupations. The root of the hierarchy will contain general biographical activities.

References

John Prager has been with IBM Research since 1979. His research interests lie within Question Answering, specifically in user and domain models, and related fields of Artificial Intelligence. Currently he is on the DeepQA research team adapting Watson, the computer system designed to play at champion human level at the Jeopardy! question-answering quiz game, for the health-care domain. John’s Ph.D. (1979) at the University of Massachusetts was in low-level vision. He received his MA (1977), Diploma in Computer Science with Distinction (1975), and BA (1974) from the University of Cambridge, UK.

Elena Filatova is an Assistant Professor at the Department of Computer and Information Sciences in Fordham University. Her research interests lie in the area of Natural Language Processing, specifically in question-answering, domain modeling, and the use of crowdsourcing for corpora generation and evaluation. Elena received her Ph.D. (2008) in Computer Science with focus on Natural Language Processing from Columbia University; MS (2001) in Computational Linguistics from USC; and Diplomas in Applied Mathematics (1998) and Philology (1999) from Moscow State University, Russia.