To be published in the proceedings of HLT/EMNLP 2005 Vancouver, Canada

Tell Me What You Do and I'll Tell You What You Are: Learning Occupation-Related Activities for Biographies

Elena Filatova*
Department of Computer Science
Columbia University
New York, NY
filatova@cs.columbia.edu

John Prager
IBM T.J. Watson
Research Center
Yorktown Heights, NY
jprager@us.ibm.com

Abstract

Biography creation requires the identification of important events in the life of the individual in question. While there are events such as birth and death that apply to everyone, most of the other activities tend to be occupation-specific. Hence, occupation gives important clues as to which activities should be included in the biography. We present techniques for automatically identifying which *important* events apply to the general population, which ones are occupation-specific, and which ones are person-specific. We use the extracted information as features for a multi-class SVM classifier, which is then used to automatically identify the occupation of a previously unseen individual. We present experiments involving 189 individuals from ten occupations, and we show that our approach accurately identifies general and occupation-specific activities and assigns unseen individuals to the correct occupations. Finally, we present evidence that our technique can lead to efficient and effective biography generation relying only on statistical techniques.

1 Introduction

Natural language processing (NLP) 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 and can give long answers to definition, biography, opinion and other question types. Such systems use general summarization techniques and at the same time take advantage of the fact that the selected information is not *generally* important but should be targeted towards answering the user's request. This idea is exploited in DUC 2004,[1] where one of the tasks is to create a summary targeting an answer to the "Who is X?" question. The systems that showed the highest performance for this task, combine traditional summarization techniques and also have modules developed specifically for creating summaries containing biographical information. Using a set of biographical facts proved to be useful to answer questions other than "What is X?" (Prager et al., 2004).

To extract biographical facts it is useful to understand the nature of different human activities. Biographical activities such as birth, death, living somewhere are applicable to all people, while each occupation is associated with its own set of activities.

In this work we suggest a novel unsupervised approach for automatic extraction of general biographical activities and activities typical for people of a particular occupation. To extract such activities we use *atomic events* (Filatova and Hatzivassiloglou, 2003) and statistical algorithms based on Markov Chains. Before creating a biography of a person as a representative of some occupation it is necessary to find out the occupation of this person; to do this we use SVM classification with activities as features.

In Section 2 we describe the current approaches for biography creation. In Section 3 we give an overview of atomic events and formulate our approach towards extracting occupation-specific activities. In Section 4 we describe mathematical models we use to identify activities typical for different occupations. In Section 5 we describe an SVM classification procedure for assigning people to the appropriate occupations. Finally, Section 6 summarizes the results and suggests a plan for future work.

*Part of this work was conducted while Elena Filatova was a summer intern at the IBM T. J. Watson Research Center.

[1]Document Understanding Conferences are a testbed for evaluating summarization systems

2 Related work

The systems participating in DUC 2004 create 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. (2004a) treat "Who is X?" as a definition question and use the DefScriber system (Blair-Goldensohn et al., 2004b). Biryukov et al. (2005) use Topic Signatures (Lin and Hovy, 2000) constructed around the person's name. Zhou et al. (2004) use nine features which are likely to be used in biography texts: bio (biographical facts), fame, personality, social, education, nationality, scandal, personal, work. Using manual annotation of 130 biographies they learn the textual patterns corresponding to these nine features.

Biographical information can be used to answer not only "Who is X?" questions. Prager et al. (2004) 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. For example, a natural constraint for artists, composers and writers is that all their works are produced in the span of time between the dates of birth and death.

Biographies vary greatly in length, genre and the presented information. Biographies of the same person in encyclopedias and yellow press magazines might contain different information. Encyclopedic biographies contain dates of birth and death, the most important achievements of people, while yellow press magazines tend to describe less important, usually scandalous facts of someone's life.

In this work we assume that most biographies can be broken into two main parts: biographical facts (the person's place and date of birth, where the person lived); and activities typically associated with the person's occupation (e.g., singers sing, explorers travel to study new lands, artists create paintings). Existing research shows that knowing the persons's occupation is helpful for detecting information which should be used in the biography (Schiffman et al., 2001; Duboue and McKeown, 2003).

3 Automatic extraction of activities typical for different occupations

3.1 Data

We created our own set of people belonging to various occupations as we were aware of no set diverse enough to analyze activities of people belonging to different occupations. We could not use the list of people whose biographies were created for DUC 2004 task: as the input documents for this task were contemporary newswire articles, more than a half of the 50 people used there were politicians.

We therefore performed what might be considered a pilot study. We chose 10 occupations and 20 practitioners of each. We understood that 10 occupations would not cover every person mentioned in a news corpus, but that was not critical to our study.

As described later, we sought documents for each chosen person. Since no documents were found for some of the individuals, these people were eliminated from the experiments; 189 survived. We ended up with the following sets of collections:

a. 20 artists	f. 20 mathematicians
b. 18 athletes	g. 19 physicists
c. 20 composers	h. 20 politicians
d. 15 dancers	i. 20 singers
e. 17 explorers	j. 20 writers

We found that for some occupations human annotators agree upon its representatives however different they are. For example, to whatever 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 the post held (president, prime-minister). Choosing an appropriate occupation title becomes crucial at the document retrieval stage as this title is used to query the search engine.

Our goals for the occupation list are that it satisfies the following criteria:

- it is diverse and covers a substantial variety of occupations from arts, sciences and other aspects of human activities;
- it contains some occupations which are closely related between each other and might be later merged into one superclass, for example, mathematicians and physicists;
- it contains occupations that are very different and it is almost impossible to specify activities which are routinely performed in two occupations, for example, singers and explorers.

To get the lists of people belonging to each particular occupation we use WordNet 2.0 (Fellbaum, 1998) (e.g., hyponyms for *composer* contain a list of composers). We also use "Google Sets" interface,[2] it was previously successfully used to find people belonging to the same occupation (Prager et al., 2004).

We retrieve documents from four corpora: AQUAINT, TREC, part of World Book and part of

[2] http://labs.google.com/sets

Encyclopedia Britannica. For document retrieval we use IBM's JuruXML search engine (Carmel et al., 2001) which allows one to index terms along with any associated named entity class labels. Queries to JuruXML may include tagged terms, which will only match similarly tagged instances in the index. We use ⟨person⟩ and ⟨role⟩ tags in the queries to perform word sense disambiguation of two types: 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 to differentiate two different people with the same name belonging to different occupations (e.g., Louis Armstrong a singer and Lance Armstrong an athlete). The second issue can be partially avoided by submitting full name of a person but it reduces the amount of documents retrieved about this person dramatically. Thus, we retrieve all the documents about people by submitting the query "⟨*person*⟩*Name*⟨*/person*⟩ ⟨*role*⟩*Occupation*⟨*/role*⟩".

The number of documents retrieved varied from one, for the query "⟨*person*⟩*Cauchy*⟨*/person*⟩ ⟨*role*⟩*mathematician*⟨*/role*⟩," to up to 8,144, for "⟨*person*⟩*Clinton*⟨*/person*⟩ ⟨*role*⟩*politician*⟨*/role*⟩." 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).

3.2 Extracting occupation-specific activities

To automatically discover general and occupation-specific activities we use a modified version of atomic events (Filatova and Hatzivassiloglou, 2003). Atomic events are triplets consisting of two named entities and a verb which labels the relation between these two named entities. We extract 189 lists of atomic events 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.
4. Count how many times each triplet containing two named entities and a verb in-between appears in the collection of documents describing the person under analysis.

First Named Entity	Verb	Second Named Entity
Columbus/PERSON	died/VBN	1506/DATE
Columbus/PERSON	sailed/VBD	India/PLACE

Table 1: A sample of atomic events extracted for the collection of documents about Christopher Columbus

First Named Entity	Verb	Second Named Entity
Vespucci/PERSON	explored/VBD	S. America/PLACE
Bering/PERSON	explored/VBD	Aleutian/PLACE

Table 2: A sample of atomic events extracted for two representatives of the explorer occupation

The NE tagger we use is a derivative of that described in (Prager et al., to appear). It tags named entities of about 100 types. Some of the marked types are very specific, like ZIPCODE and ROYALTY. To avoid overfitting we choose six high-level types for atomic events' extraction: PERSON, PLACE, DATE, WHOLENO, ORG and ROLE. In contrast to the original atomic event scores we keep simple counts for the triplets as later we combine triplets extracted for different people. Table 1 contains examples from the list of atomic events extracted for Columbus.

3.3 Generalized atomic events

Our goal is to collect information about activities general for all people and about activities specific for some occupations. Thus, we are interested in the semantic information reflected in the atomic events but not in the exact named entities. We analyze not the atomic events themselves but the generalized versions of the extracted atomic events. For example, here are two sentences about explorers:

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

The corresponding atomic events extracted for these sentences are presented in Table 2. Clearly, these atomic events capture information about the same type of activity, namely that explorers explore some locations. What makes these atomic events different is the exact names of the explorers and the locations a particular explorer explored. We can unify these atomic events by omitting the exact named entities and leaving only their types. The resulting atomic events we call *generalized atomic events*. The atomic events presented in Table 2 can be converged to the following generalized atomic event:

> NAME/PERSON - explored/VBD - PLACE

In the generalized atomic events we distinguish two types of named entities with the tag PERSON: those which refer to the person under analysis (from now

on they are marked as NAME/PERSON) and all the rest (marked as PERSON). Thus, we separate the person whose occupation we want to identify from the people who are linked to this person through some activities. This generalization technique is similar to the one used by Yangarber(2003) for semantic patterns discovery for information extraction.

Filatova and Hatzivassiloglou (2003) showed that atomic events capture the most important relations described in the input and assign to them good-quality labels. In this work we show that generalized atomic events can be used for capturing the activities performed by people of different occupations.

To select occupation-related activities we merge lists of atomic events corresponding to the people of the same occupation. Hence, we get ten lists of generalized atomic events corresponding to the ten occupations under analysis. The count of each generalized atomic event is equal to the sum of the counts of all the atomic events which are merged into this generalized atomic event.

4 Getting 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 (pre-Google) by Kleinberg (1998) 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 discovering activities closely related to a specific occupation to the problem of ranking the reliability of web-pages for the submitted query.

To rank the importance of activities for a particular occupation we define a bipartite graph $G = \{N, V, E\}$, where V are the verb nodes (activities), N are the nodes corresponding to the named entity types linked to V verbs, and E are the arcs connecting the named entity types and the activities. A part of such a bipartite graph created for the *explorers* occupation is presented in Figure 1.

Following this procedure we create bipartite

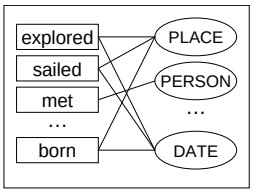

Figure 1: Bipartite graph for a set of generalized atomic events corresponding to *explorers* occupation

Dancer	Physicist	Singer
made/VBD	born/VBN	said/VBD
died/VBD	died/VBD	born/VBN
appeared/VBD	announced/VBD	died/VBD
been/VBN	discovered/VBD	join/VB
founded/VBD	be/VB	singing/VBG
became/VBD	including/VBG	sang/VBD
born/VBN	became/VBD	has/VBZ
danced/VBD	wrote/VBD	conducting/VBG
blessed/VBN	helped/VBD	made/VBD
perform/VB	named/VBN	became/VBD

Table 3: Top ten activities for three occupations: dancers, physicists and singers

graphs for every occupation. Each of the created ten bipartite graphs contains m named entity types on one side and k verbs (activities) on the other side.[3]

We define $P_{N\to V}$ as a $m \times k$ stochastic transition matrix from named entities to verbs, with elements[4]

$$P_{N\to V}[i,j] \equiv p_{n_i,v_j} = (1-c)\frac{f(n_i \to v_j)}{\sum_{v\in V} f(n_i \to v)} + c \quad (1)$$

where f is equal to the sum of the counts of all the generalized atomic events containing this link for the occupation under analysis. In the same way we define $P_{V\to N}$ $k \times m$ row-stochastic transition matrix from verbs to named entities. Using $P_{V\to N}$ and $P_{N\to V}$, we can define the transition matrix:

$$P_{V\to V} = P_{V\to N} \cdot P_{N\to V} \quad (2)$$

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 stochastic.According to Markov Chain Theory (Kemeny and Snell, 1960) 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 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 (2) ranks how tightly the activities

[3] Variables m and k are unique for every occupation

[4] To avoid data sparseness we use a smoothing factor $c = 0.01$.

Biography-related verbs
said/VBD
born/VBN
died/VBD
wrote/VBD
became/VBD
had/VBD
known/VBN
be/VB
included/VBD
including/VBG

Table 4: Top ten activities common for all the eleven occupations.

are linked to the occupation under consideration. The size of this matrix depends on the variety of the verbs in all forms used in the generalized atomic events for the representatives of this occupation and varies from 800 for physicists up to 2100 for politicians in our data.

Table 3 contains top ten activities for three occupations: 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 eigenvectors corresponding to the eigenvalues of 1.

The activities presented in Table 3 can be divided into three types:

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

For our classification we rely mainly on the first type of the activities. To extract the activities of the second type we create $P_{V \to V}$ a transition matrix for the combined set for all the generalized atomic events created for all the ten occupations. 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 4 contains top ten activities for this matrix.

In Section 5 we show that the lists of occupation-related and general activities are reliable features for classifying people according to their occupations.

5 Classification

In this section we describe people classification according to their occupations. For our classification experiments we use a multi-class SVM classifier.[5] 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-specific activities (Section 5.1); the other set consists of the complete triplets for the generalized atomic events (Section 5.2).

5.1 SVM classification using only verbs

To get verb features for multi-class SVM classification we use ten occupation-related lists of activities with activities sorted according to the eigenvector corresponding to the steady state.

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 top 15% of the activities from each of the ten occupation-specific lists and the list of general activities.
- V4 From the ten occupation-specific 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 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 from occupation to occupation (for example, 1794 activities are used in the atomic events for composers, and 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, 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.

[5] http://www.cs.cornell.edu/People/tj/svm_light/svm_multiclass.html

			SVM classification				Probing		Random
			Verbs		Atomic Events				
Occupation	Amount of Representatives	Average Amount of Documents	Amount	Ratio	Amount	Ratio	Amount	Ratio	
Artists	20	10.0	9	0.450	15	0.750	14	0.700	0.106
Athletes	18	10.0	12	0.667	16	0.889	14	0.778	0.095
Composers	20	9.65	10	0.500	15	0.750	19	0.950	0.106
Dancers	15	9.07	7	0.467	13	0.867	11	0.733	0.079
Explorers	17	9.0	12	0.706	15	0.882	15	0.882	0.090
Mathematicians	20	7.2	10	0.500	8	0.381	20	1.000	0.106
Physicists	19	7.05	5	0.263	6	0.316	13	0.684	0.101
Politicians	20	10.0	9	0.450	12	0.600	1	0.050	0.106
Singers	20	9.05	9	0.450	12	0.600	10	0.500	0.106
Writers	20	10.0	7	0.350	12	0.600	10	0.500	0.106
Average				0.480		0.663		0.677	

Table 5: Performance of different classification methods.

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 events, then this feature receives the value of 1, if there is no atomic event using this activity as a connector then this feature is assigned 0. We use binary values for our features instead of the atomic event counts because the reliability of the scores for the atomic events 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 3documents which only mentione the people queried.

Removal of some of the features is reinforced by the (Koller and Sahami, 1997) 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.

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 events 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 a small number of the documents retrieved for these people. Table 5 shows how many documents are analyzed per person on average for each occupation. Due to the nature of our document collections, the smallest number of documents analyzed was for mathematicians and physicists: current newswire texts do not contain much information about scientists, and those parts of encyclopedias 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. As the distribution of people across occupations is not even, Table 5 contains two numbers for each occupation: the absolute number and the ratio of people classified correctly for this occupation.

We believe that the performance of SVM classification based solely on the activities is so 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 only for the explorers this activity is linked to the PLACE named entity tag. Though the classification based solely on verbs gives quite poor results we consider it to be a valid starting classification as usually activities are associated with the verbs corresponding to these activities.

5.2 SVM classification using atomic events

To create generalized atomic event features for multiclass 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.

AE1 Same as step V1 above.

AE2 Same as step V2 above.

AE3 For the top 15% of the activities from the ten occupation-specific lists get all the generalized atomic events containing those activities.

AE4 For the top 15% of the activities typical for all the occupation (Step AE2) get all the generalized atomic events containing those activities.

AE5 From the ten occupation-related lists (Step AE3) remove those generalized atomic events which are also present in the list of general generalized atomic events (Step AE4).

AE6 Merge the ten occupation-related lists into one and remove from it all the generalized atomic events that appear in more than 2 occupations.

Out of 189 people, nine are not assigned any features. This again is because all the atomic events extracted for these 9 people were either too general or too specific. The people who do not get any event features are the same as those who do not get any verb features plus one physicist. Out of the remaining 180 people, 124 people are classified into the appropriate occupations. Table 5 shows that generalized atomic events are more reliable for occupation classification than plain activities extracted from these generalized atomic events. Thus, it can be concluded that structured information captured by atomic events is valuable and reliable. For example, using atomic events Matisse was correctly classified as an artist. According to the t-test the performance of the classification based on atomic events 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 while classifying him as a composer is a valid results as well.

5.3 Other types of classification

The task of classifying people according to their occupations is new and to our knowledge there is no existing baseline we could compare our results with. Nevertheless, we decided to adapt for comparison two classification techniques used for other tasks: random assignment of an occupation and probing.

Random occupation assignment. As the distribution of people among the occupations is not even we cannot give one exact probability of assigning a correct occupation to a particular person. Instead, we calculate such random probabilities for each occupation. The results are presented in Table 5.

Random assignment gives a very low baseline which we easily outperform, which is why we use another classification based on probing to estimate how good our results are. Classification based on probing is considered to be the state-of-the-art classification method for such tasks as hidden web classification (Ipeirotis et al., 2003) and answer verification (Magnini et al., 2002).

Probing. 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.

$$Occupation_j = max_{for\ all\ i,j} \frac{count_{occupation_i,\ person_j}}{count_{occupation_i}} \quad (3)$$

According to Table 5 SVM, classification based on atomic events outperforms probing classification for six occupations out of ten, for one occupation (*explorers*) the results for SVM classification and probing are the same and for three occupations probing classification outperforms SVM classification. One of the cases where probing 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 events has performance comparable to probing-based classification. This is significant since in those tasks for which it has been used so far, probing classification outperforms other methods and is considered to be the state-of-the-art (Ipeirotis et al., 2003; Magnini et al., 2002).

5.4 Using classification extracted features

Though we do not dramatically outperform probing, our methodology has one crucial advantage. We use classification not as a primary task but as an evaluation testbed to show that the lists of generalized atomic events created for every occupation and for

Artists	Athletes
NAME - painted/VBN - DATE	NAME - win/VB - WHOLENO
NAME - resemble/VB - PERSON	NAME - scored/VBD - WHOLENO
PERSON - designed/VBN - NAME	NAME - winning/VBG - DATE
Composers	Dancers
NAME - composed/VBN - PERSON	NAME - danced/VBN - ORG
NAME - include/VBP - WHOLENO	PLACE - presented/VBD - NAME
ROLE - hearing/VBG - NAME	NAME - appeared/VBD - WHOLENO
Explorers	Mathematicians
NAME - annexes/VBZ - PLACE	PERSON - developed/VBD - NAME
NAME - reach/VB - PLACE	ROLE - prove/VB - NAME
NAME - declares/VBZ - DATE	NAME - studied/VBD - WHOLENO
Physicists	Politicians
DATE - described/VBD - NAME	NAME assassinated/VBN - PLACE
ROLE - predicted/VBD - NAME	NAME - postponed/VBN - PLACE
NAME - continued/VBD - ORG	NAME - flown/VBN - ORG
Singers	Writers
NAME - conducting/VBG - PLACE	PLACE - leaving/VBG - NAME
NAME - sing/VB - ROLE	NAME - translated/VBN - PERSON
NAME - sang/VBD - PERSON	WHOLENO - written/VBN - NAME

Table 6: Occupation-specific generalized atomic events (NAME stands for NAME/PERSON).

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 events 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 6, on the other hand, contains occupation-specific generalized atomic events. These generalized atomic events 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.

6 Conclusions and future work

We reported results on extracting human activities which can be used for classifying people according to their occupations. We introduced a novel representation for describing human activities (generalized atomic events). SVM classification using generalized atomic events as features gives results comparable to other state-of-the-art classification techniques. We are currently looking at ways of identifying other types of activities which are neither general no occupation-specific but rather person-specific. We observed that those generalized atomic events which have high scores for a particular person but are not used in the description of any other person are good candidates to point out person-specific information. We believe that the usage of generalized atomic events can enable significant new techniques for a number of natural language processing tasks. One direction is to use the generalized atomic events typical for all the occupations as an initial representation for the auxiliary biography-related questions. Another direction is to use generalized atomic events for biography generation.

References

M. Biryukov, R. Angheluta, and M.-F. Moens. 2005. Multi-document question answering text summarization using topic signatures. *Journal on Digital Information Management*, 3(1):27–33.

S. Blair-Goldensohn, D. Evans, V. Hatzivassiloglou, K. McKeown, A. Nenkova, R. Passonneau, B. Schiffman, A. Schlaikjer, A. Siddharthan, and S. Siegelman. 2004a. Columbia university at DUC 2004. In *Proceedings of DUC*, Boston.

S. Blair-Goldensohn, K. McKeown, and A. Schlaikjer, 2004b. *Answering Definitional Questions: A Hybrid Approach*, pages 47–58. AAAI Press.

D. Carmel, E. Amitay, M. Hersovici, Y. Maarek, Y. Petruschka, and A. Soffer. 2001. Juru at TREC 10. Experiments with index pruning. In *Proceedings of TREC*.

P. Duboue and K. McKeown. 2003. Statistical acquisition of content selection rules for natural language generation. In *Proceedings of the EMNLP Conference*, pages 121–128, Sapporo, Japan.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT press.

E. Filatova and V. Hatzivassiloglou. 2003. Domain-independent detection, extraction, and labeling of atomic events. In *Proceedings of the RANLP Conference*, pages 145–152, Bulgaria.

P. Ipeirotis, L. Gravano, and M. Sahami. 2003. QProber: A system for automatic classification of hidden-web resources. *ACM Transactions on Information Systems*, 21(1):1–41.

J. Kemeny and J. Snell. 1960. *Finite Markov Chains*. Princeton, NJ: Van Nostrand.

J. Kleinberg. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677.

D. Koller and M. Sahami. 1997. Hierarchically classifying documents using very few words. In *Proceedings of ICML*, pages 170–178, Nashville, US.

C.-Y. Lin and E. Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the COLING Conference*, Germany, July.

B. Magnini, M. Negri, R. Prevete, and H. Tanev. 2002. Is it the right answer? Exploiting web redundancy for answer validation. In *Proceedings of the 40th Annual ACL Meeting*, pages 425–432, Philadelphia, USA.

J. Prager, J. Chu-Carroll, and K. Czuba. 2004. Question answering using constraint satisfaction: QA-by-Dossier-with--Constraints. In *Proceedings of the 42nd Annual ACL Meeting*, pages 575–582, Spain.

J. Prager, J. Chu-Carroll, E. Brown, and K. Czuba, to appear. *Question Answering by Predictive Annotation*. Kluwer Academic Publishers.

B. Schiffman, I. Mani, and K. Concepcion. 2001. Producing biographical summaries: combining linguistic knowledge with corpus statistics. In *Proceedings of the 39th Annual ACL Meeting*, pages 450–457, Toulouse, France.

R. Yangarber. 2003. Counter-training in discovery of semantic patterns. In *Proceedings of the ACL Conference*, Japan.

L. Zhou, M. Ticrea, and E. Hovy. 2004. Multi-document biography summarization. In *Proceedings of the EMNLP Conference*, pages 434–441, Spain.